2026年 フロンティアAIモデルの評価と展望：DeepSeek V4 Proの包括的分析と「8ヶ月の遅れ」が示唆する業界の地殻変動

Gemini 3.1 pro

1. イントロダクション：2026年第2四半期におけるAI開発の転換点

2026年4月は、大規模言語モデル（LLM）および推論モデル（LRM）の開発競争において歴史的な転換点として記憶されることとなる。米国を拠点とするOpenAIが次世代フロンティアモデル「GPT-5.5」を、Anthropicが「Claude Opus 4.7」を相次いでリリースする中、中国のAI開発企業であるDeepSeekは、オープンウェイトモデルの最前線を再定義する「DeepSeek V4 Pro（以下、DeepSeek V4）」および軽量版の「DeepSeek V4 Flash」を公開した[1]。このリリースは、単なるモデル規模の拡大にとどまらず、エージェント型AIワークフローの基盤となる「コンテキスト長」と「推論コスト」の力学を根本から覆すものである[4]。

DeepSeek V4のプレビュー版リリースは、OpenAIのGPT-5.5発表から24時間以内に行われた[5]。このタイミングは、オープンソースのフロンティアモデルがプロプライエタリ（非公開）なAIの進化のペースに単に追随するだけでなく、そのコスト構造と効率性において業界のペースメーカーとなりつつあることを市場に示す明確な意思表示であった[5]。しかし、モデルの能力を測る評価基準（ベンチマーク）が飽和状態にある現在、開発企業による自己申告のスコアと、実際の運用環境や非公開の検証環境におけるパフォーマンスとの間には、無視できない乖離が生じている[6]。

本報告書は、米国標準技術研究所（NIST）傘下の人工知能標準化・イノベーションセンター（CAISI）が2026年5月1日に発表したDeepSeek V4の独立評価レポート[6]を中核に据え、複数の独立系ベンチマーク（BenchLM、Chatbot Arenaなど）やオープンソースコミュニティの実証データ[8]を統合的に分析する。特に、CAISIの評価が指摘した「フロンティア（最先端）モデルからの8ヶ月の遅延」という結論の真意を解き明かし[6]、自己申告ベンチマークでは米国最高峰のモデル（GPT-5.4やClaude Opus 4.6）に匹敵するとされるDeepSeek V4が、なぜクリーンな評価環境では数世代前の水準に留まるのかを多角的に考察する[6]。さらに、本報告では、アーキテクチャの革新、ソフトウェアエンジニアリング（コーディング）における実用性、中国国内のエコシステムにおける位置づけ、コスト効率、そして安全性における新たな脆弱性について、網羅的かつ深い洞察を提供する。

2. CAISI評価の詳細分析：能力評価における「8ヶ月の遅れ」の真意

モデルの自己申告スコアと、外部の独立機関による評価の間には、しばしば解釈の余地が生じる。DeepSeekの開発データによれば、DeepSeek V4は2026年2月～3月頃にリリースされた「Claude

Opus 4.6」や「GPT-5.4」と同等の能力を有するとされている[6]。事実、SWE-Bench Verifiedなどの公開ベンチマークでは、DeepSeek V4 Proが80.6%〜81%というスコアを叩き出し、Claude Opus 4.6の80.8%に肉薄している[6]。しかし、米国標準技術研究所（NIST）傘下のCAISIが実施した徹底的な評価は、全く異なる景色を提示している。

2.1 汚染されていないベンチマークによる真の実力測定

CAISIは、サイバーセキュリティ、ソフトウェアエンジニアリング、自然科学、抽象的推論、数学の5つのドメインにわたって、9つのベンチマークを用いてDeepSeek V4を評価した[6]。この評価の核心は、学習データセットに含まれている可能性のある公開ベンチマーク（データ汚染のリスクがあるもの）だけでなく、外部に公開されていない「内部構築ベンチマーク」を使用した点にある。

CAISIが独自に構築したソフトウェアエンジニアリング評価「PortBench」と、抽象的推論を測る半公開データセット「ARC-AGI-2 semi-private」を用いたテストにおいて、DeepSeek V4のスコアは、フロンティアモデルの水準を下回った[6]。具体的には、DeepSeek V4の能力は、2025年8月にリリースされた「GPT-5」とほぼ同等であるとCAISIは結論付けている[6]。GPT-5からGPT-5.4、そしてGPT-5.5やClaude Opus 4.6へと至る開発のタイムラインを考慮すると、DeepSeek V4の基礎的な推論能力は、米国の最先端モデルに対して「約8ヶ月遅れている」ことになる[6]。

CAISIはモデルの能力を項目応答理論（Item Response Theory: IRT）に触発された手法で適合させ、16のベンチマークと35のモデルを用いて集約的な能力比較図を作成した[6]。この分析において、Y軸の200ポイントの上昇は、特定のタスクを解決するオッズが3倍に増加することを意味するが、DeepSeek V4のトレンドラインは明確に米国モデルの後塵を拝している[6]。

2.2 領域別の能力格差とデータ汚染の疑義

CAISIの評価結果は、公開データセットと非公開データセットの間におけるパフォーマンスの乖離を如実に示している。以下の表は、CAISIが実施した評価におけるDeepSeek V4 ProとGPT-5.5のスコア比較である。

評価ドメイン	ベンチマーク名	**DeepSeek V4 Pro**	**GPT-5.5 (xhigh)**	特記事項
ソフトウェアエンジニアリング	SWE-Bench Verified	81%	81%	公開データセット
ソフトウェアエンジニアリング	PortBench	78%	78%	CAISI内部構築（非公開）

自然科学	FrontierScience	79%	-	-
専門的知識	GPQA-Diamond	96%	-	大学院レベルの推論
抽象的推論	ARC-AGI-2 semi-private	79%	-	半公開データセット
サイバーセキュリティ	CTF-Archive-Diamond	71%	71%	-

データソース: CAISI Evaluation Report (May 2026) [6]

公開ベンチマークであるSWE-Bench VerifiedにおいてGPT-5.5と同等の81%を記録しながら、非公開のPortBenchでは78%に留まるという事実は極めて重要である[6]。この3%の低下は、モデルがパブリックなコーディングタスクに過学習（オーバーフィット）しているか、あるいは学習データに公開テストセットが意図せず混入している（データ汚染）可能性を強く示唆している[6]。

さらに、GPT-5.5 (xhigh) が同CAISIの評価においてSWE-bench Verifiedで81%、PortBenchで78%、CTF-Archive-Diamondで71%という同一または近似したスコア構成を示している点も特筆すべきである[6]。この一致は、DeepSeek V4がGPT-5.5の特定の振る舞いを模倣（蒸留）しているか、あるいは最先端モデルが直面する現在の技術的限界（頭打ち）に達していることを示している。外交問題評議会（CFR）の分析においても、DeepSeekを含む中国企業が米国モデルに対する「非合法な蒸留攻撃（illicit distillation attacks）」によってデータを生成し、米国モデルの能力の一部を低コストで複製しているという疑惑が指摘されている[16]。

これらのデータは、DeepSeek V4 Proが「汎用的な知能」において米国を追い抜いたわけではないことを明確に示している。しかし、モデルの真の価値は単一の知能スコアだけで決まるものではない。次節以降で詳述する通り、このモデルが業界に与えた真の衝撃は、その計算効率とアーキテクチャの革新にある。

3. アーキテクチャと技術的革新：100万トークン時代の計算効率化

DeepSeek V4の最大の功績は、純粋な知能の絶対値を引き上げたことではなく、極端に長いコンテキスト（100万トークン）を処理する際の限界費用を劇的に引き下げた点にある[4]。エージェント型AIが

自律的にタスクを遂行するためには、過去のツールの実行結果や数百ファイルに及ぶコードベースをすべてプロンプトに保持し続ける必要があり、従来のアーキテクチャではKVキャッシュ（Key-Value Cache）の枯渇やアテンション計算の爆発的な増大という壁に直面していた[4]。

3.1 パラメータ構成とMixture-of-Experts (MoE) 設計

DeepSeek V4 Proは、総パラメータ数1.6兆（1.6T）のMoE（Mixture-of-Experts）アーキテクチャを採用している[2]。しかし、推論時にアクティブになるパラメータ数は約490億（49B）に抑えられており、各トークンの処理においてモデル全体のごく一部の「専門家（Expert）」ネットワークのみが活性化される設計となっている[7]。一方、軽量版のDeepSeek V4 Flashは、総パラメータ数2,840億（284B）、アクティブパラメータ数130億（13B）であり、より高速かつ経済的な推論を目的としている[2]。

この極端なMoE設計は、理論上の処理効率を飛躍的に高める一方で、本質的なルーティングのオーバーヘッドや推論パイプラインの複雑化を招く[7]。1.6兆パラメータという数字は知識の巨大なデータベースとして機能するが、単一の推論ステップにおいてその全てが動員されるわけではない。この「疎（Sparse）」な活性化メカニズムが、後述するベンチマークにおけるパフォーマンスのムラの一因となっている[7]。

3.2 ハイブリッドアテンション（CSA + HCA）と極限の効率化

DeepSeek V4の技術的飛躍の中心には、新たなアテンション機構の導入がある。本モデルは「Compressed Sparse Attention（CSA）」と「DeepSeek Sparse Attention（DSA）」、およびスライディングウィンドウを組み合わせたハイブリッド・アテンションアーキテクチャ（CSA + HCA）を採用している[2]。CSAは、学習された圧縮重みを介して m 個のKVトークンごとに1つのエントリに圧縮し、その後Lightning Indexerを介したトップ k 選択によるスパースアテンションを適用する[17]。この仕組みは、トークン単位の圧縮を行い、超長文コンテキストの推論コストを劇的に削減する。

具体的には、100万トークンのコンテキスト入力時において、DeepSeek V4 Proの1トークンあたりの推論FLOPs（浮動小数点演算数）は、前世代のDeepSeek V3.2と比較してわずか27%にまで削減されている[4]。さらに重要な点として、GPUのメモリ帯域を圧迫する最大の要因であるKVキャッシュの消費量が、V3.2のわずか10%に抑えられている[4]。V4 Flashに至っては、FLOPsが10%、KVキャッシュが7%という驚異的な効率化を達成している[4]。

また、学習プロセスにおいて「Manifold-Constrained Hyper-Connections（mHC）」と呼ばれる技術を導入し、深い残差アーキテクチャにおける信号の伝播を表現力を犠牲にすることなく安定化させている[17]。32兆〜33兆トークンという膨大な事前学習データセットに対する最適化には、スケールアップされたMuon Optimizerが使用された[17]。

計算資源の最適化という観点からは、MoEの専門家パラメータにはFP4（4ビット浮動小数点数）精度を用い、その他の主要なパラメータにはFP8精度を用いるという「混合精度（Mixed Precision）トレーニング」が採用されている[19]。これにより、モデルのパフォーマンスを妥協することなく、メモリ効率を最大化している。

4. エージェント機能とソフトウェアエンジニアリングの実力

現代のLLMにおける最大の戦場は、自律的にコードを生成・修正し、ターミナルを操作する「エージェント的ソフトウェアエンジニアリング」の領域である。DeepSeek V4 Proは、オープンウェイトモデルとして過去最高レベルの到達点を示しているが、タスクの性質によって得意・不得意が明確に分かれる。

4.1 リポジトリレベルの推論とコーディングベンチマーク

ソフトウェアエンジニアリングの能力を測定する標準的な指標として定着したSWE-Benchにおいて、DeepSeek V4 Proは特筆すべき成果を上げている。厳密な検証が行われた「SWE-Bench Verified」において、DeepSeek V4 Proは80.6%〜81%を記録し、Claude Opus 4.6（80.8%）とほぼ同等の性能を示した[1]。

このスコアの背後にある実用的な能力は、内部やリークされたベンチマークテストにおいても確認されている。DeepSeek V4は、単一機能の生成において92.10%の成功率を誇り、GPT-5.4 Miniの87.40%を上回る[23]。複数ファイルのバグ修正においても83.70%の解決率を記録し（同GPT-5.4 Miniは33.80%）、アーキテクチャの大規模なリファクタリングにおいても78.50%の成功率を示した[23]。これにより、生成されたパッチが既存のCI/CDパイプラインのテストを初回でパスする確率が極めて高くなっている。

ベンチマーク指標	**DeepSeek V4 Pro**	**GPT-5.4**	**Claude Opus 4.6**	**Claude Opus 4.7**
Codeforces Rating	3206	3168	該当データなし	該当データなし
SWE-Bench Verified	80.6% - 81%	~80.0% (未検証含)	80.8%	87.6%
SWE-Bench Pro	55.4%	57.7%	74.0%	64.3%
LiveCodeBench	93.5%	該当データなし	88.8%	該当データなし

Terminal-Bench 2.0	67.9%	75.1%	65.4%	69.4%

データソース: 独立系ベンチマークおよび公式リリースデータ [11]

上記の表が示す通り、最新のプログラミングコンテストの課題を用いるLiveCodeBenchにおいてDeepSeek V4 Proは93.5%を記録し、Claude Opus 4.6（88.8%）を凌駕してオープンソースのSOTA（State-of-the-Art）を獲得している[12]。競技プログラミングのプラットフォームであるCodeforcesのレーティングにおいても3206を達成し、GPT-5.4（3168）やGemini-3.1-Pro Highを上回っている[13]。

4.2 ターミナル操作と視覚的推論における遅れ

一方で、シェルスクリプトの実行、ファイルシステムのナビゲーション、ビルドツールの操作など、開発者の日常的なターミナル操作を自律的にシミュレートする「Terminal-Bench 2.0」においては、異なる様相を呈する。このベンチマークではGPT-5.4が75.1%という圧倒的なスコアで首位に立ち、DeepSeek V4 Proは67.9%に留まった[13]。Claude Opus 4.6は65.4%であり、DeepSeek V4 ProはClaudeを上回るものの、複雑なCLI（コマンドラインインターフェース）ワークフローや、自己修復を伴う長期的なエージェントループにおいては、依然としてOpenAIのアーキテクチャが優位性を保っている[24]。GPT-5.4は「Interactive Thinking」と呼ばれる推論途中のプラン調整機能を導入しており、複雑なエラーハンドリングにおいて強みを発揮する[25]。

視覚的推論（マルチモーダル機能）においても同様の傾向が見られる。画像、チャート、アーキテクチャ図を解析する「MMMU Pro」において、GPT-5.4はツールなしで81.2%を記録しているのに対し、Claude Opus 4.6はツールありでも77.3%に留まる[24]。DeepSeek V4 Proは純粋なテキストおよびコードの処理に特化しており、画像入力のネイティブサポートを持たないため[14]、視覚情報を含むフルスタックなフロントエンド開発やUIデバッグにおいては、GPT-5.5やClaude Opus 4.7といった最先端モデルに大きく遅れをとっている[27]。

5. コンテキスト長と数学推論：MoEアーキテクチャの光と影

DeepSeek V4のアーキテクチャは長大なコンテキストを極めて低コストで処理できるが、その能力が全ての領域で均等に向上しているわけではない。特に数学的推論において、その限界が観察されている。

5.1 100万トークンウィンドウの精度と「Engram」メモリ

DeepSeek V4 ProとFlashは、ともに100万（1M）トークンのコンテキストウィンドウを標準でサポートしている[2]。これは、単なる「容量」の確保にとどまらず、実用的な情報検索精度を伴っている。「Engram」と呼ばれる条件付きメモリ技術により、1MトークンのNeedle-in-Haystack（干し草の山から針を探す）テストにおいて97%という高い検索精度を達成している[15]。また、学術的な長文コンテキストベンチマークであるMRCR 1Mにおいても83.5%を記録し、Gemini 3.1 Proを上回る堅牢なパフォー

マンスを示した[20]。

比較として、Claude Opus 4.6もベータ版として1Mコンテキストを提供しているが、極端な長さでの検索品質に関する公開データは少ない[15]。GPT-5.4はネイティブで272Kのウィンドウを持ち（Codex構成で最大1M[1]）、128Kを超えるプロンプトには追加のサーチャージが発生する[15]。この点において、巨大なリポジトリ全体を一度にプロンプトへ流し込むようなユースケースにおいて、DeepSeek V4は独自の優位性を持っている。

5.2 数学推論における回帰現象（Regression）

しかし、DeepSeek V4 Proの評価において専門家が注視している懸念点の一つが、一部の推論タスク、特に数学的ベンチマークにおいて観察される「回帰（Regression）」である。高度な数学タスクを測るCMathにおいて、前世代のDeepSeek V3.2からV4への移行時にスコアのわずかな低下（Pro版で92.6%から90.9%への低下など）が見られ、MGSM（多言語小学校算数）などのベンチマークでもパフォーマンスの安定性にムラがあることが報告されている[7]。MATH-500でのスコアも約88.3%であり、Claude Opus 4.7の89.4%に僅差で後れを取っている[29]。

この現象は、コンテキストの長さ（100万トークン）を処理する能力が向上した一方で、その長大なコンテキストを「深く理解し、推論を貫徹する能力」が比例して向上していないことを示唆している[7]。コンテキストウィンドウの拡大は単なる「容量」の増加であり、知識の結びつけや論理的なジャンプの精度を必ずしも保証しない。また、1.6兆パラメータに対してわずか490億のパラメータしかアクティブにならない巨大なMoEルーティングに依存するアーキテクチャでは、推論プロセスが複数のExpertネットワークに分散される[7]。この「重いルーティングのオーバーヘッド」により、特定の数学的論理展開において文脈の喪失やエラーの蓄積が起きやすくなっており、エレガントな設計というよりも「巨大なモデルを動かすためのサバイバル・エンジニアリング」の産物であるという厳しい指摘も存在する[7]。

6. コスト効率と破壊的価格設定：マルチモデル・ルーティング時代の幕開け

DeepSeek V4 Proが市場に与えた最も深刻な衝撃は、ベンチマークのスコアではなく、その「価格破壊」にある。LLMを製品に組み込む開発者や、自律型マルチエージェントシステムを運用する企業にとって、推論コストはモデルの知能と同等かそれ以上に重要な決定要因である[1]。

6.1 フロンティアモデルとの圧倒的な価格差

2026年4月時点での主要なフロンティアモデルのAPI価格（100万トークンあたり）を比較すると、DeepSeek V4の優位性は圧倒的である。

モデル名	入力価格（**1M**トークン）	出力価格（**1M**トークン）	コンテキスト長	知能インデックス

DeepSeek V4 Flash	~$0.14	~$0.28	1M	-
DeepSeek V4 Pro (Max)	$1.74	$3.48	1M	52
GPT-5.4	$2.50	$15.00	1.05M	-
Claude Sonnet 4.6 (Max)	$6.56	-	1M	52
Claude Opus 4.6	$15.00	$75.00	1M	-
GPT-5.5 (xhigh)	$11.25	-	922K	60

データソース: Artificial Analysis Leaderboards および各種公式ドキュメント [15]

Claude Opus 4.6と比較すると、DeepSeek V4 ProのAPI価格は入力トークンで約8分の1（実勢価格の比較研究によっては約50倍のコスト差と評価されるケースもある）、出力トークンでは実に約21分の1（$3.48 vs $75.00）に設定されている[12]。例えば、CI/CDパイプラインに統合されたコードベース解析エージェントが1日に1000万トークンを処理する場合、年間コストはGPT-5.4で約40,000ドル、Claude Opus 4.6で約58,000ドルに達するが、DeepSeek V4 Proを使用すれば約1,400ドルに収まるという試算がある[15]。

CAISIの評価においても、DeepSeek V4は同様の能力を持つ他のモデルと比較して極めて費用対効果が高いことが立証されている。米国で最もコスト競争力のあるリファレンスモデル（GPT-5.4 mini）と比較した場合でも、7つのベンチマークのうち5つでV4が優れたコスト効率を示し、全体として53%安価から41%割高というレンジに収まっている[6]。

6.2 スペシャリスト・ルーティング・アーキテクチャの標準化

この破壊的な価格設定は、「推論を複数回ループさせる」マルチエージェント・アーキテクチャの進化を加速させている。単一のモデルにすべてのタスクを依存することは、技術的負債として毎月複利で

蓄積されるリスクとなっている[1]。

AI.ccのようなモデルルーティングAPIを活用し、各モデルを最も得意とする領域に割り当てる「スペシャリスト・ルーティング・アーキテクチャ」が2026年の標準的な設計思想となりつつある[1]。例えば、バルク処理、大量のドキュメント検索、日常的なコーディングにはDeepSeek V4 FlashやProを割り当て、極めて複雑な意思決定、長時間の自律的コーディングループ、マルチモーダル（画像や動画）解析が必要な局面のみGPT-5.5やClaude Opus 4.7にルーティングするというハイブリッド戦略である[1]。

7. 実運用における摩擦：「Thinking」ループの罠とインフラ要件

DeepSeek V4 ProはAPI価格において圧倒的な魅力を放つが、実環境（特にエージェント・ハーネスやIDE）にデプロイする際、開発者は「オープンウェイト特有の摩擦」とも呼ぶべき技術的課題に直面する。

7.1 推論状態の不一致：HTTP 400エラー問題

DeepSeek V4は、高精度の推論を行うために「Thinkingモード（Chain-of-Thought）」を内蔵している。このモードを有効にしてAPIを呼び出すと、モデルは最終的な回答（content）とは別に、内部の思考プロセスを記録したreasoning_contentという追加フィールドを生成して返す[32]。問題は、DeepSeekのAPI仕様が、同一の会話コンテキスト内で後続のプロンプトを送信する際、過去のターンのreasoning_contentをそのままAPIに送り返すことを要求する点にある[32]。

現在普及しているLangGraph、Cursor、Hermes-agent、OpenClawなどの多くのエージェントループや開発環境は、AnthropicやOpenAIのAPI仕様（メッセージフォーマット）に最適化されており、この独自のreasoning_contentフィールドを保持・中継するようには設計されていない[33]。結果として、複数ターンの複雑なエージェント作業を行わせようとすると、APIからHTTP 400エラー（"reasoning_content must be passed back to the API"）が返され、ワークフローが途中でクラッシュする事態が頻発している[33]。これを回避するためには、思考ブロックをコンテキストから削除して再入力するハックや、Thinkingモード自体をオフにする必要があり、モデル本来の推論能力を制限することに繋がっている[32]。

7.2 セルフホスティングのハードルと地政学的制約

さらに、自社インフラでセルフホストする場合、1.6兆パラメータというサイズは、FP8精度であってもエンタープライズ級のGPUクラスタを要求する。小規模なラップトップや単一GPUのワークステーションでは到底運用できず、最低でも8基のハイエンドGPUノードが必要となる[37]。MoE特有のルーティング・オーバーヘッドにより、バッチサイズが小さい環境ではスループットが低下しやすく、応答時間（レイテンシ）は必ずしも低くない。実際、Artificial Analysisのデータによれば、GPT-5.5 (high) の応答時間が28.53秒であるのに対し、DeepSeek V4 Pro (Max) は144.34秒に達しており、出力速度も34トークン/秒と相対的に遅い[30]。

また、DeepSeek V4は、NVIDIAのGPUではなく、中国の通信機器大手Huawei（ファーウェイ）のAIアクセラレータ「Ascend 910B」ファミリでの推論に最適化されて開発・拡張されたと報告されている[16]。

これは、米国によるH20などの先進AIチップの輸出規制（2025年4月）に対する適応の結果である[38]。しかし、V4の技術レポートにおいて、モデルの学習に具体的にどのチップが使用されたかが意図的に伏せられている点は特筆に値し、中国のAI開発が依然として米国技術のエコシステム（クラウド経由の演算リソース利用や蒸留）から完全に独立できていないことを示唆している[16]。

8. 中国のオープンウェイトAIエコシステムにおける圧倒的優位性

米中間のAI技術競争というマクロな視点で見ると、DeepSeek V4 Proの登場は、中国国内のAIエコシステムにおける決定的なリードを確立したことを意味する。CAISIのレポートが明記している通り、DeepSeek V4は「これまでに評価された中で最も能力の高いPRC（中華人民共和国）モデル」である[6]。

独立系評価プラットフォームであるBenchLMのデータ（2026年5月1日更新）を参照すると、中国製AIモデルの勢力図が明確になる。DeepSeek V4 Pro (Max) は全体スコア88/100を獲得し、中国モデルのリーダーボードで首位に立っている[10]。

モデル名	開発元	プロビジョナルスコア	コンテキスト長	ライセンス形態	特記事項
DeepSeek V4 Pro (Max)	DeepSeek	88	1M	オープンウェイト	コーディング89.8、中国首位
Kimi K2.6	Moonshot AI	84	256K	オープンウェイト	サブエージェント並列処理に強み
GLM-5.1	Z.AI	83	203K	オープンウェイト	高い知識スコア(85.1)、MITライセンス
GLM-5 (Reasoning)	Z.AI	82	200K	オープンウェイト	数学・推論に特化

Qwen 3.6 Plus	Alibaba	74	1M	プロプライエタリ	エージェント操作・多言語に強み

データソース: BenchLM Chinese Models Directory (May 2026) [10]

Z.AIが開発した「GLM-5.1」は全体スコア83で4位につけている。203Kのコンテキストウィンドウを持つオープンウェイトモデルであり、汎用知識（GPQA等）や指示追従性（IFBench等）において極めて高いパフォーマンスを発揮する[10]。特に数学的推論において強みを持つが、純粋なコーディング性能においてはDeepSeek V4に一歩譲る[10]。しかし、MITライセンスでの提供により、エンタープライズのファインチューニング用途で根強い支持を集めている[41]。

Alibabaの「Qwen 3.6 Plus」は全体スコア74であり、DeepSeek V4と同様に100万トークンのコンテキストをサポートする。オープンウェイトではなくプロプライエタリ（API経由のみ）で提供されているが、エージェント的コーディングにおいて極めて堅牢であり、複雑なツール使用において高い評価を得ている[10]。LMSYSのChatbot Arenaにおけるコーディング Eloスコアでは、Claude Opus 4.6が1549でグローバル首位を独占する中[8]、Qwen 3.6 Plusは1504、GLM-5.1は1519を記録しており、中国勢が米国トップ層に肉薄していることが伺える[10]。

DeepSeek V4 Proは、これらの競合モデルをコーディング性能とコスト効率の両面で圧倒しており、事実上、中国国内および世界のオープンウェイト界隈における「ベースライン・インフラストラクチャ」としての地位を確立しつつある。

9. セキュリティと安全性：推論モデル（LRM）特有の脆弱性

強力な推論能力（Chain-of-Thought）を備えた次世代モデル（Large Reasoning Models: LRMs）の台頭は、AIアライメントとセキュリティの分野に新たなパラダイムをもたらしている。CAISIによるレッドチーム評価（脆弱性テスト）や学術機関の大規模な実証研究により、LRM特有の脆弱性が明らかになっている[42]。

9.1 CoTアタックによるジェイルブレイクの進化

従来のLLMに対するジェイルブレイク（安全フィルターの回避）は、複雑なロールプレイや仮想のシナリオ（「これは架空の物語である」等の指示）をプロンプトに埋め込む手法が主流であった[44]。しかし、DeepSeek V4 ProのようなLRMにおいては、「推論プロセス（思考ブロック）」そのものをハッキングする手法が有効であることが判明している。

ある大規模な実証研究では、32のオープンソースモデルとDeepSeek V3.2/V4、Gemini 3 Pro、GPT-5.2を含む8つの商用モデルに対して、56種類のジェイルブレイク技術と4種類の「CoT攻撃（Chain-of-Thought Attack）」を実行し、計460万回のAPIコールによる評価を行った[43]。その結果、ユーザー側から特定のレスポンスプレフィックス（応答の冒頭部分）を強制するCoT攻撃を用いることで、攻撃成功率が平均で3.4倍に跳ね上がり、一部のモデルでは成功率が0.6%から96.5%へと劇的

に悪化することが示された[43]。さらに、セキュリティ会社Penligentのレポートによれば、公開されている特定のインスタンスにおいては、フィルタリングされていないジェイルブレイクの成功率が100%に達したケースも確認されている[23]。

これは、モデルに「悪意のある行動を正当化する論理的な思考プロセス」を強制的に開始させると、モデル自身が持つ強力な推論能力が、安全フィルターのロジックを自ら論破・無効化する方向に働いてしまうという逆説的な現象である[43]。推論能力が高ければ高いほど、一度誤った（あるいは悪意のある）前提に立脚して思考を始めた場合、それを強固に自己正当化してしまうリスクを孕んでいる。

9.2 安全性アライメントの劣化とエンタープライズ需要

さらに同研究は、ポストトレーニング（強化学習や知識蒸留）のプロセスにおいて、一般的な能力（汎用知能）の向上を過度に優先するあまり、安全性アライメントが体系的に劣化する傾向があることを指摘している[43]。DeepSeek V4 Proは「Answer-Then-Check（回答してから確認する）」という後段の安全性チェック機構を取り入れることで防衛を図っているが、防衛成功率（DSR）の観点からは依然として課題が残る[45]。

厳密なコンプライアンスや予測可能な挙動が求められる金融機関や医療機関のエンタープライズワークフローにおいては、AnthropicのClaude Opus 4.6や4.7が提供する強固なセーフティガードレールとSLA（サービスレベルアグリーメント）が依然として不可欠であると評価されている[12]。安全性がビジネスリスクに直結する領域では、単なるコスト効率以上に、モデルの拒絶行動の予測可能性が重要視されるためである[12]。

10. 結論：DeepSeek V4 Proの歴史的意義と「8ヶ月の壁」の未来

本報告書の多角的な分析を通じて、DeepSeek V4 Proに対する包括的な評価は次のように総括される。

1. 経済性の再定義とアーキテクチャの勝利:
   DeepSeek V4 Proは、100万トークンの長大なコンテキストを実用的なコストで処理する技術（ハイブリッドアテンションと極端なMoE設計）において、間違いなく業界のブレイクスルーを達成した。これにより、エージェントAIの設計パラダイムは根本から変化し、より複雑で反復的なプロンプトチェーンが経済的に実行可能となった。
2. 純粋な推論能力における「8ヶ月の壁」:
   自己申告のベンチマークや、データ汚染の疑いがある公開テストにおいて、DeepSeek V4 Proは米国の最先端モデル（GPT-5.4やClaude Opus 4.6）と互角に渡り合っているように見える。しかし、CAISIが実施した厳格かつ非公開の評価環境（PortBench等）においては、そのパフォーマンスは2025年8月にリリースされたGPT-5のレベルに留まっており、真の汎用的・未知の推論能力においては依然として約8ヶ月の遅れが存在する。
3. 実運用における課題とインフラ税:
   オープンウェイトモデルとしての可用性は高いものの、reasoning_contentのループ要件など、既存のエージェント・インフラとの統合には固有の摩擦が存在する。また、複雑なCLI操作や視

覚的推論においては米国フロンティアモデルのエコシステムに及ばず、自社ホスティングには膨大なハードウェア投資が必要となる。

4. セキュリティの新たな地平:
   LRM（Large Reasoning Models）という新たなカテゴリに属する本モデルは、高度な推論能力を持つがゆえに、CoT攻撃によるジェイルブレイクに対して極めて脆弱であるという新たなセキュリティ上の課題を浮き彫りにした。

結論として、DeepSeek V4 Proは「絶対的な知能」において米国フロンティアモデルを完全に追い抜いたわけではない。しかし、「GPT-5クラスの知能の限界費用を限りなくゼロに近づけた」という点において、AI業界全体のソフトウェア経済を根底から作り変える推進力（Driving Force）となった。企業や開発者は、最高水準の推論と堅牢性を要するタスクには引き続きGPT-5.5やClaude 4.7を採用しつつ、スケーラビリティとコスト効率が支配する広範なタスクをDeepSeek V4に委譲するという、ハイブリッドなマルチモデル戦略を採用することが、2026年以降の最も合理的な選択となるだろう。

引用文献

1. From GPT-5.5 to DeepSeek V4: How Developers Are Building Smarter AI Agents with Multi-Model Routing in 2026 - AiThority, 5月 3, 2026にアクセス、https://aithority.com/machine-learning/from-gpt-5-5-to-deepseek-v4-how-developers-are-building-smarter-ai-agents-with-multi-model-routing-in-2026/
2. DeepSeek V4 Preview Release, 5月 3, 2026にアクセス、https://api-docs.deepseek.com/news/news260424
3. Agentic AI Models: The Latest Developments and News, 5月 3, 2026にアクセス、https://www.crescendo.ai/blog/agentic-ai-models
4. DeepSeek-V4: a million-token context that agents can actually use, 5月 3, 2026にアクセス、https://huggingface.co/blog/deepseekv4
5. DeepSeek V4 Preview: The Complete 2026 Guide - o-mega | AI, 5月 3, 2026にアクセス、https://o-mega.ai/articles/deepseek-v4-preview-the-complete-2026-guide
6. CAISI Evaluation of DeepSeek V4 Pro - National Institute of Standards and Technology, 5月 3, 2026にアクセス、https://www.nist.gov/news-events/news/2026/05/caisi-evaluation-deepseek-v4-pro
7. DeepSeek V4 is Shitty - Medium, 5月 3, 2026にアクセス、https://medium.com/data-science-in-your-pocket/deepseek-v4-is-shitty-b067af243019
8. LMSYS Chatbot Arena Coding Leaderboard April 2026: The New Superintelligence Tier, 5月 3, 2026にアクセス、https://aidevdayindia.org/blogs/lmsys-chatbot-arena-current-rankings/lmsys-chatbot-arena-coding-leaderboard-2026.html
9. DeepSeek V4 Pro Benchmarks 2026: Scores, Rankings & Performance | BenchLM.ai, 5月 3, 2026にアクセス、https://benchlm.ai/models/deepseek-v4-pro
10. Best Chinese LLMs in 2026: DeepSeek V4, Kimi K2.6, GLM-5, Qwen ..., 5月 3, 2026にアクセス、https://benchlm.ai/blog/posts/best-chinese-llm
11. LLM Leaderboard 2026 — Compare Top AI Models - Vellum, 5月 3, 2026にアクセス、https://www.vellum.ai/llm-leaderboard

12. DeepSeek V4 Alters Everything We Knew About Price-Performance Math - Lightning AI, 5月 3, 2026にアクセス、https://lightning.ai/blog/deepseekv4comparison
13. DeepSeek V4 Pro Review: Benchmarks, Pricing & Performance (2026) - Codersera, 5月 3, 2026にアクセス、https://codersera.com/blog/deepseek-v4-pro-review-benchmarks-pricing-2026/
14. DeepSeek V4 Pro (Reasoning, Max Effort) vs GPT-5 (high): Model Comparison, 5月 3, 2026にアクセス、https://artificialanalysis.ai/models/comparisons/deepseek-v4-pro-vs-gpt-5
15. DeepSeek V4 vs Claude Opus 4.6 vs GPT-5.4: AI Coding Model Comparison (2026), 5月 3, 2026にアクセス、https://www.nxcode.io/resources/news/deepseek-v4-vs-claude-opus-vs-gpt-5-coding-2026
16. DeepSeek V4 Signals a New Phase in the U.S.-China AI Rivalry, 5月 3, 2026にアクセス、https://www.cfr.org/articles/deepseek-v4-signals-a-new-phase-in-the-u-s-china-ai-rivalry
17. deepseek-ai/DeepSeek-V4-Pro · Technical Report Summary - Hugging Face, 5月 3, 2026にアクセス、https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/discussions/129
18. DeepSeek's new models offer big inference cost savings, 5月 3, 2026にアクセス、https://www.theregister.com/2026/04/24/deepseek_v4/
19. deepseek-ai/DeepSeek-V4-Pro - Hugging Face, 5月 3, 2026にアクセス、https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro
20. DeepSeek V4 Pro: Model Overview, Features & Performance Guide - DeepInfra, 5月 3, 2026にアクセス、https://deepinfra.com/blog/deepseek-v4-pro-model-overview
21. DeepSeek-V4-Pro-Max Benchmarks, Pricing & Context Window - LLM Stats, 5月 3, 2026にアクセス、https://llm-stats.com/models/deepseek-v4-pro-max
22. Claude Opus 4.6 vs GPT-5.4: Full Benchmark Breakdown (2026) | BenchLM.ai, 5月 3, 2026にアクセス、https://benchlm.ai/blog/posts/claude-opus-vs-gpt-5
23. DeepSeek V4 SWE-bench Score: The Ultimate Guide to the New AI Coding Frontier, 5月 3, 2026にアクセス、https://skywork.ai/skypage/en/deepseek-v4-ai-coding-guide/2047581548426514433
24. GPT-5.4 vs Claude Opus 4.6: a guide to choosing the right model - Portkey, 5月 3, 2026にアクセス、https://portkey.ai/blog/gpt-5-4-vs-claude-opus-4-6/
25. Claude Opus 4.7 vs. GPT-5.4: Which Frontier Model Should You Use? - DataCamp, 5月 3, 2026にアクセス、https://www.datacamp.com/blog/opus-4-7-vs-gpt-5-4
26. GPT-5.5 vs DeepSeek V4: Benchmarks, Pricing and Which to Use | DataCamp, 5月 3, 2026にアクセス、https://www.datacamp.com/blog/deepseek-v4-vs-gpt-5-5
27. GPT-5.5 VS Deepseek V4 Pro VS Opus 4.7: I tested THEM on My KingBench 2.0 Questions!, 5月 3, 2026にアクセス、https://www.youtube.com/watch?v=exS-Y6XGk6s
28. DeepSeek V4: Features, Benchmarks, and Comparisons - DataCamp, 5月 3, 2026

にアクセス、https://www.datacamp.com/blog/deepseek-v4
29. DeepSeek V4: The Open-Source Model That Rivals Closed Frontier Models | MindStudio, 5月 3, 2026にアクセス、https://www.mindstudio.ai/blog/deepseek-v4-open-source-frontier-model-review
30. LLM Leaderboard - Comparison of over 100 AI models from OpenAI ..., 5月 3, 2026にアクセス、https://artificialanalysis.ai/leaderboards/models
31. DeepSeek V4 vs GPT-5.4 vs Claude Opus 4.6 (April 2026) - EvoLink.AI, 5月 3, 2026にアクセス、https://evolink.ai/blog/deepseek-v4-vs-gpt-5-4-vs-claude-opus-4-6-verified-comparison
32. DeepSeek V4 Pro + OpenClaw: Why does the second message always fail with reasoning_content must be passed back? - Reddit, 5月 3, 2026にアクセス、https://www.reddit.com/r/openclaw/comments/1subtbk/deepseek_v4_pro_openclaw_why_does_the_second/
33. DeepSeek V4: context limited to 200K + reasoning_content error - Bug Reports - Cursor, 5月 3, 2026にアクセス、https://forum.cursor.com/t/deepseek-v4-context-limited-to-200k-reasoning-content-error/159045
34. [BUG] DeepSeek v4-pro reasoning mode breaks with "reasoning_content must be passed back" error · Issue #16135 · NousResearch/hermes-agent - GitHub, 5月 3, 2026にアクセス、https://github.com/NousResearch/hermes-agent/issues/16135
35. DeepSeek /anthropic (V4 thinking): stripped thinking blocks cause HTTP 400 on replay · Issue #16748 · NousResearch/hermes-agent - GitHub, 5月 3, 2026にアクセス、https://github.com/NousResearch/hermes-agent/issues/16748
36. DeepSeek V3.2 looping bug: what settings / harness tweaks are actually reducing it in production? : r/AI_Agents - Reddit, 5月 3, 2026にアクセス、https://www.reddit.com/r/AI_Agents/comments/1swcp29/deepseek_v32_looping_bug_what_settings_harness/
37. DeepSeek V4 Pro for Local Vulnerability Discovery, What Actually Works - Penligent, 5月 3, 2026にアクセス、https://www.penligent.ai/hackinglabs/deepseek-v4-pro-for-local-vulnerability-discovery-what-actually-works/
38. OSINT Report: DeepSeek V4 release timeline, internal training bottlenecks, and the shift from Huawei to NVIDIA. April 2026 Prediction. - Reddit, 5月 3, 2026にアクセス、https://www.reddit.com/r/DeepSeek/comments/1s694qu/osint_report_deepseek_v4_release_timeline/
39. DeepSeek V4 Pro (Reasoning, Max Effort) vs GLM-5.1 (Reasoning): Model Comparison, 5月 3, 2026にアクセス、https://artificialanalysis.ai/models/comparisons/deepseek-v4-pro-vs-glm-5-1
40. Which OpenCode Go model is your favorite? : r/opencodeCLI - Reddit, 5月 3, 2026にアクセス、https://www.reddit.com/r/opencodeCLI/comments/1swlnu9/which_opencode_go_model_is_your_favorite/

41. The Best Open-Source LLMs for Agentic Coding in 2026 - MindStudio, 5月 3, 2026にアクセス、https://www.mindstudio.ai/blog/best-open-source-llms-agentic-coding-2026
42. Center for AI Standards and Innovation (CAISI) | NIST, 5月 3, 2026にアクセス、https://www.nist.gov/caisi
43. What Matters For Safety Alignment? - arXiv, 5月 3, 2026にアクセス、https://arxiv.org/html/2601.03868v2
44. Has Deepseek v4 become positively aligned compared to previous versions? - Reddit, 5月 3, 2026にアクセス、https://www.reddit.com/r/SillyTavernAI/comments/1sw3yua/has_deepseek_v4_become_positively_aligned/
45. Reasoned Safety Alignment: Ensuring Jailbreak Defense via Answer-Then-Check, 5月 3, 2026にアクセス、https://openreview.net/forum?id=DK6AToxJNo