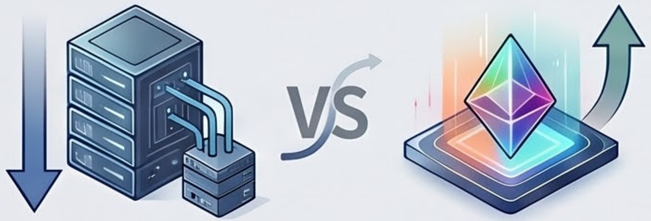
2026年中国次世代LLMの衝撃：自律型エージェントと圧倒的コスト効率が拓く新潮流

パラダイムシフト：2026年のAI評価軸



パラメーター規模の拡大　　効率と自律性

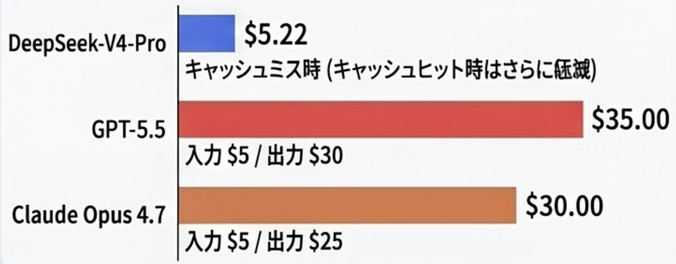
「規模」から「効率と自律性」への移行

パラメータ数の競争は終結し、推論時のコンピュート効率と、長時間にわたる自律的なタスク完遂能力（長期オーケストレーション）が新たな評価指標となりました。

「金魚の記憶」問題の克服

従来のモデルが長い対話の中で文脈を失う問題を、アーキテクチャの再設計（KVキャッシュ圧縮や思考の保存）により克服しています。

最先端モデルのAPI利用コスト（100万トークンあたり）



DeepSeek-V4：限界費用の破壊者

米国製最高峰モデルの約1/6のコスト

API価格を劇的に引き下げ、100万入力・出力トークンの合計コストを5.22ドルに抑えることで、「エージェント経済学」を再定義しました。

KVキャッシュのサイズを98%削減

ハイブリッド圧縮アテンション（HCA）の採用により、標準的なアーキテクチャと比較してKVキャッシュを約2%まで圧縮し、100万トークンの実用化を達成。

Moonshot Kimi K2.6：スウォーム知能の極致

最大300のサブエージェントによる並列実行

単一モデルの強化（スケールアップ）ではなく、タスクを分解し多数のエージェントで分担する「スケールアウト」により、4,000ステップの自律作業を並列化。

13時間の自律セッションでコードを全面改修

金融マッチングエンジンの改修において、4,000行以上のコードを変更し、スループットを185%向上させるという長期間発能力を実証。

Alibaba Qwen3.6：計算の線形化と思考の保持

Gated DeltaNetによるO(n)の計算効率

文脈長が伸びると計算量が指数関数的に増える弱点を、線形アテンション機構の導入により解決し、101万トークンの超長文脈を効率処理。

思考の保存（Thinking Preservation）

対話履歴から接続プロセスを保持して次ターンへ引き続ぐことで、エージェントが同じエラーを繰り返す「ループによる失敗」を防止。

Xiaomi MiMo-V2.5-Pro：IoTとのネイティブ統合

「人・車・家」を繋ぐ自律の脳

HyperOSと融合し、ユーザーのカレンダーや家中のスマートデバイスの状況をリアルタイムで理解し、自律的に家電やアクションを制御。

動画理解ベンチマークでGemini 3 Proに肉薄

Video-MMEで87.7を記録し、数分間の映像からシーンや時系列を正確に理解するフロンティア級のマルチモーダル能力を保持。

総合性能比較（ベンチマーク）

指標	DeepSeek-V4-Pro	Kimi K2.6	MiMo-V2.5-Pro	Claude 4.6 Opus (参考)
SWE-Bench Verified	80.6%	80.2%	-	80.8%
Terminal-Bench 2.0	67.9%	66.7%	43.2%	65.4%
最大文脈長	100万	26万	100万	-