

Moonshot AI 「Kimi K2 Thinking」 モデル：詳細調査レポート

調査日: 2025年11月12日

対象モデル: Kimi K2 Thinking (月之暗面 / Moonshot AI)

エグゼクティブサマリー

重要な結論:

- Kimi K2 Thinkingは2025年11月6日にリリースされた推論特化型オープンソースAIモデルで、トレーニングコストは約460万ドル、Nvidia H800 GPUで訓練された
- GPT-5やClaude Sonnet 4.5といったプロプライエタリモデルを、エージェント推論などの主要ベンチマークで上回る初のオープンウェイトモデル
- 1兆パラメータのMixture-of-Experts (MoE) アーキテクチャで、推論時は320億パラメータのみが活性化
- 200~300回の連続ツール呼び出しを人間の介入なしに実行可能な、長期的エージェント機能を実現
- INT4量子化を用いて推論速度を約2倍に向上させ、256Kトークンのコンテキストウィンドウをサポート

AIコミュニティの反応: AI業界では「オープンソースモデルがこれまでで最もクローズドなフロンティアモデルに接近した」と評価され、DeepSeek R1のo1への追随と同様の転換点とされている

1. Moonshot AI（月之暗面）企業概要

1.1 設立と経営陣

Moonshot AIは2023年3月に清華大学交叉情報研究院出身の楊植麟(Yang Zhilin)、周昕宇(Zhou Xinyu)、呉育昕(Wu Yuxin)によって北京で設立された。社名はPink Floydのアルバム『The Dark Side of the Moon』の発売50周年に因んで命名され、これは創業者楊植麟の最も好きなアルバムである。

1.2 資金調達と企業評価額

設立後約2年で約12.7億ドルの資金を2回の資金調達ラウンドで獲得し、企業評価額は33億ドルに達した。主要な投資家にはAlibaba、Tencent、Meituan、HongShan（旧Sequoia China）などが含まれる。2025年10月には、IDG Capitalが主導する約6億ドルの新たな資金調達ラウンドを完了間近であると報じられている。

1.3 企業ビジョンと戦略

楊植麟CEOは、AGI（汎用人工知能）達成を目指す基盤モデルの構築を目標に掲げ、「損失のない長コンテキスト」機能の開発、AGI構築、エンタープライズではなく消費者向けアプリケーションの作成という3つの重点分野に注力している。「次の時代において、OpenAIの技術的理想主義とByteDanceの商業哲学を組み合わせた企業になることを望む」と表明している。

1.4 製品開発の歴史

Kimiチャットボットの進化:

- 2023年10月: 初代Kimiチャットボットをリリース、20万中国語文字の処理が可能
- 2024年3月: Kimiが200万中国語文字を単一プロンプトで処理可能に大幅アップグレード
- 2025年1月20日: Kimi K1.5をリリース、OpenAI o1と同等の数学・コーディング・マルチモーダル推論性能を主張

Kimi K2シリーズ:

- 2025年7月: Kimi K2（1兆パラメータの大規模言語モデル）のウェイトを公開、MoEアーキテクチャで推論時に320億パラメータが活性化、15.5兆トークンで訓練、Modified MITライセンスで公開
- 2025年9月9日: Kimi-K2-Instruct-0905をリリース、エージェントコーディングタスクの性能を向上させ、コンテキストウィンドウを128Kから256Kトークンに倍増
- 2025年11月: Kimi K2 Thinkingをリリース、高度な推論とエージェントタスク向けのオープンソースアップデート、約460万ドルで訓練

2. Kimi K2 Thinkingモデルの技術仕様

2.1 アーキテクチャの詳細

Kimi K2 Thinkingは、1兆の総パラメータを持つMixture-of-Experts (MoE)モデルで、推論時に320億パラメータが活性化する。

主要な技術仕様:

- レイヤー数:** 61層（Dense層を含む）、そのうちDense層は1層
- 注意機構の隠れ次元:** 7,168
- エキスパート数:** 384個のエキスパート、トークンごとに8個を選択
- 共有エキスパート:** 1個
- 語彙サイズ:** 160K
- コンテキスト長:** 256K
- 注意機構:** MLA (Multi-Head Latent Attention)
- 活性化関数:** SwiGLU

2.2 主要な技術的特徴

1. 深い思考とツールオーケストレーション Chain-of-Thought推論と関数呼び出しをインターリーブするエンドツーエンド訓練により、数百ステップにわたるドリフトなしの自律的な研究、コーディング、執筆ワークフローを実現。

2. ネイティブINT4量子化 ポストトレーニング段階で量子化認識トレーニング(QAT)を採用し、MoEコンポーネントにINT4ウェイトオンリー量子化を適用することで、低レイテンシモードで損失のない2倍のスピードアップを実現。すべてのベンチマーク結果はINT4精度で報告されている。

- 3. 安定した長期エージェント機能** 最大200〜300回の連続ツール呼び出しにわたって一貫した目標指向の動作を維持し、30〜50ステップ後に劣化する従来のモデルを超越。
- 4. テストタイムスケーリング** 推論トークン数とツール使用のステップ数の両方を増やすテストタイムスケーリングに取り組み、これがK2 Thinkingの進歩を支える技術的方向性である。

2.3 技術的イノベーション

Muonオプティマイザー Kimi K2は、AI研究者Keller Jordanらが開発したMuonオプティマイザーの実装における真のイノベーションを実現した。Moonshot AIとUCLAの研究者は2025年初頭に、MuonがLLMトレーニングにスケーラブルであることを示す研究論文を発表。従来のオプティマイザーは支配的な方向に固執する傾向があるが、Muonは行列の直交化を使用してこれを解決し、モデルがより広い解空間を探索し、局所的なトラップを回避できるようにする。

3. ベンチマーク性能と競合比較

3.1 エージェント推論タスク

HLE (Humanity's Last Exam) テキストのみ (ツールあり) において、K2 Thinkingは44.9%を達成し、GPT-5の41.7%、Claude Sonnet 4.5 Thinkingの32.0%を上回った。

BrowseCompベンチマーク (エージェント検索とブラウジングをテスト) では、K2 Thinkingが60.2%を記録し、GPT-5の54.9%、Claude Sonnet 4.5の24.1%を大きく上回った。

3.2 数学的推論

AIME25 (ツールなし) では94.5%、Pythonツールありでは99.1%を達成し、GPT-5 (ツールなし94.6%、ツールあり99.6%) とほぼ同等の性能。

HMMT25では、ツールなしで89.4%、Pythonツールありで95.1%を達成。

3.3 コーディングタスク

SWE-bench Verified (ツールあり) では71.3%を記録し、K2 0905の69.2%を上回ったが、GPT-5の74.9%、Claude Sonnet 4.5の77.2%には及ばなかった。

SWE-bench Multilingual (ツールあり) では61.1%を達成し、GPT-5の55.3%、K2 0905の55.9%を上回り、Claude Sonnet 4.5の68.0%には及ばなかった。

LiveCodeBench (公式テスト) では約83%の精度を達成し、DeepSeek-V3.2の約74%を大きく上回った。

3.4 一般タスク

GPQA Diamondでは84.5%を記録し、GPT-5の85.7%、Grok-4の87.5%にわずかに及ばなかった。

MMLU-Proでは84.6%、MMLU-Reduxでは94.4%を達成。

3.5 競合モデルとの比較

DeepSeek-V3.2との比較: K2 Thinkingは、同じくオープンソースの競合モデルMiniMax-M2を大幅に上回った。BrowseComp結果60.2%はM2の44.0%を超え、SWE-Bench Verified 71.3%はM2の69.4%をわずかに上回った。

技術的差異: 両モデルともスパースMixture-of-Expertsアーキテクチャを採用しているが、Moonshotのネットワークはより多くのエキスパートを活性化し、高度な量子化認識トレーニング(INT4 QAT)を展開している。

3.6 Heavy Modeの性能

K2 Thinking Heavy Modeは効率的な並列戦略を採用し、まず8つの軌跡を同時にロールアウトし、次にすべての出力を反動的に集約して最終結果を生成する。Heavy ModeでのHLE（ツールあり）では51.0%を達成。

4. 市場への影響とインパクト

4.1 オープンソースAIの転換点

Kimi K2 Thinkingの登場は、オープンソースとクローズドモデルが高性能領域で収束する転換点を示す構造的シフトである。かつてプロプライエタリAPIに独占的に依存していた企業は、GPT-5レベルの推論と同等のオープン代替品を展開でき、ウェイト、データ、コンプライアンスの完全な制御を維持できる。

4.2 コスト効率と経済的影響

トレーニングコスト: CNBCの報道によれば、内部関係者の情報として、このモデルのトレーニングコストはわずか460万ドルで、DeepSeek V3のレンタル価格560万ドル、R1の29.4万ドルと比較される。

推論コスト: Moonshot AIは、Kimi K2 Thinkingの価格を、キャッシュヒット時は100万トークンあたり0.15ドル、キャッシュミス時は0.60ドル、出力は2.50ドルとしている。これらのレートは、MiniMax-M2の入力0.30ドル、出力1.20ドルと比較しても競争力があり、入力1.25ドル、出力10ドルのGPT-5と比較すると1桁低い。

4.3 中国AI産業の地位向上

年初にはAIに少し詳しい人でも中国のAI研究所をおそらく知らなかったが、2025年末に向けて、DeepSeek、Qwen、Kimiがすべて著名な名前になっている。中国企業は、DeepSeek R1のリリース後わずか6ヶ月で「オープンフロンティアの性能レベル」に追いついた。

Stability AI CEOは、開発者が先進的なBlackwellチップにアクセスできれば、最先端の性能達成コストが300万ドル未満になる可能性があるかと推測している。

4.4 地政学的インパクト

Nathan Lambertは「年初にAIを緩く追っている人々の大半は中国のAI研究所を0社知っていたと思われるが、現在、2025年末に向けて、DeepSeek、Qwen、Kimiはすべて著名な名前になっている」と述べた。中国

企業がDeepSeekの初期リリース後わずか6ヶ月で「オープンフロンティアの大まかな性能」に追いついたことを強調し、従来の西洋ハブ外での加速的なイノベーションのペースを浮き彫りにした。

4.5 プロプライエタリモデルへの圧力

独立した評価で、K2のような自由に利用可能なモデルが、困難な試験でトップティアのプロプライエタリモデルを上回るスコアを示すとき、最も先進的なAIを構築できるのは技術大手だけという概念が覆される。クローズドモデルプロバイダーには、純粋な性能の観点からは競争の場が平準化されているため、より高い信頼性、安全性、統合を通じてプレミアムオフリングを正当化する明確な圧力がかかっている。

4.6 市場での反応と評価

初期段階のベンチャーキャピタルファームMenlo VenturesのパートナーDeedy Dasは、金曜日のXでの投稿で「今日はAIの転換点だ。中国のオープンソースモデルが1位になった」と書いた。

Kimiのサーバーはすでに完全に圧倒されていると報告されている。

5. 実用的な応用と展開

5.1 利用可能性とアクセス

アクセス方法:

- kimi.comのチャットモードおよびKimi K2 Thinking APIを通じて利用可能
- Modified MITライセンスの下、Hugging Faceで正式にリリース
- platform.moonshot.aiでAPIにアクセス可能

ライセンス条項: 標準MITライセンシングの自由を保持しつつ、1つの制限を追加: ソフトウェアまたは派生製品が月間アクティブユーザー数1億人以上、または月間収益2,000万ドル以上のサービスを提供する場合、展開者は製品のユーザーインターフェースに「Kimi K2」を目立つように表示する必要がある。

5.2 展開と推論エンジン

Kimi-K2-Thinkingは、vLLM、SGLang、KTransformersなどの推論エンジンでの実行が推奨される。

愛好家たちはXで、2台のMac Studio (M3 Ultra)システムでKimi K2 Thinkingを実行でき、15トークン/秒の速度で3,500トークンを生成できると共有している。

5.3 実用的な応用例

複雑なアプリケーション生成: Moonshotによって共有された例には、単一のプロンプトからWordクローンやウイルスシミュレーターなどの複雑なアプリケーションを生成するKimi K2 Thinkingが含まれている。

多段階問題解決: モデル作成者によれば、ある例では、K2 Thinkingが23回の絡み合った推論とツール呼び出しを通じて博士レベルの数学問題を解決した。

5.4 制約と課題

冗長性: この高性能には重大なトレードオフがある: 冗長性。モデルは非常に長い出力を生成し、Artificial Analysis評価スイートを完了するために1億4,000万トークンの合計を使用する。これは、DeepSeek V3.2などの競合他社が使用するトークンの約2.5倍である。このような高い冗長性は、クエリの実行コストと完全な応答を受け取るのにかかる時間の両方を増加させることにより、低いトークンあたりの価格の利点の一部を相殺し、実用的な応用に直接影響する。

モダリティの制限: 最先端モデルがマルチモーダルを標準機能としている今日の時代において、Kimi K2 Thinkingは依然として純粋なテキストモデルである。

6. 今後の展望と業界への影響

6.1 オープンソースAIの加速

Moonshot AIのオープン公開戦略は、DeepSeek R1、Qwen3、GLM-4.6、MiniMax-M2によって設定された先例に従うが、それを完全なエージェント推論に拡張している。学術および企業開発者にとって、K2 Thinkingは透明性と相互運用性の両方を提供し、推論トレースを検査し、ドメイン固有のエージェントのパフォーマンスを微調整する能力を提供する。

6.2 競争環境の変化

Kimi K2 Thinkingのリリースは、中国の競争の激しいAI市場における計算された動きである。かつての新星であったMoonshot AIは、DeepSeek、Z.ai、Alibabaなどの国内ライバルから激しい圧力に直面している。この新モデルは、競争の焦点をスピードと価格から洗練された推論とエージェントパフォーマンスに移すことで、リーダーシップを取り戻す直接的な試みである。

6.3 エージェントAIの未来

MoonshotのK2 Thinkingによる最新の取り組みは、AI業界における根本的なグローバルシフトの一部である。企業は、単にテキストを生成したりコードを提案したりするチャットボットを超えて進化している。新たなフロンティアはエージェント知能: 高レベルの目標を理解し、計画を策定し、それを実行するためにさまざまなデジタルツールを使用できる自律システムの作成である。

6.4 エネルギー効率と持続可能性

モデル間の究極の競争は、最終的にエネルギーをめぐる競争になる。中国のオープンソースモデルはエネルギー消費において大きな利点を持ち、中国自体がエネルギー資源において巨大な利点を持っている。

6.5 技術開発の民主化

これらの日々、多くの人々が優れたAIモデルを訓練することを非常に迅速に学んでいることは、見落とされがちな刺激的なことの1つである。最先端のAIモデルを訓練し、国際的に配布する能力は、グローバルに普及することになる。

7. 技術的課題と今後の改善点

7.1 現在の制限事項

チャットモードの制限: 高速で軽量の体験を確保するため、kimi.comのチャットモードでは選択的にツールのサブセットを採用し、ツール呼び出しステップの数を削減している。その結果、kimi.comでのチャットではベンチマークスコアを再現できない可能性がある。完全な機能を反映するAgenticモードは近日中に更新される予定。

7.2 今後の改善領域

マルチモーダル機能の追加: 現時点では純粋なテキストモデルであるため、視覚や空間推論を含むタスクを処理する際に制限がある。

コーディング性能の向上: SWE-benchでは競争力があるものの、GPT-5やClaude Sonnet 4.5にはまだ若干及ばない。

冗長性の最適化: 現在の高い冗長性（1億4,000万トークン）を削減し、実用的なコスト効率を改善する必要がある。

8. AI分野の将来への影響分析

8.1 オープンソース vs プロプライエタリモデルの力関係変化

Kimi K2 Thinkingの登場により、オープンソースモデルとプロプライエタリモデルの性能差が事実上消滅しつつある。オープンソースが勝利したのは、それが運命づけられていたからではなく、同等の性能で月額24ドルと240ドルの選択肢があれば、24ドルが毎回勝つからである。

8.2 エージェント技術の進化

K2 Thinkingモデルのリリースに関して人々が話題にしていることの1つは、Kimi K2 Thinkingがクエリに答える際に「数百のツール呼び出し」を使用することである。これは、従来のLLMから自律的なAIエージェントへの進化を示す重要な転換点である。

8.3 中国AI産業のグローバル影響力

最先端のマインドシェアの増大する割合が中国にシフトしている。Z.ai、Meituan、Ant Lingなどが来年このリストに加わる可能性がある。これらの研究所の中には、文字通りDeepSeek R1の後に基盤モデルの取り組みを開始したものもある。

8.4 AI開発のコスト構造の変革

未確認の数字によれば、Kimi K2は460万ドルの予算で訓練されたとされている（ただし、DeepSeek-R1で示されたように、これが全体像ではない可能性がある）。最先端に近い性能がもはや少数の重資金投入された西

洋研究所の独占ではなくなり、プロプライエタリモデルへの圧力が強まり、生のAI知能が商品化する可能性を示唆している。

8.5 推論モデルの競争激化

Gemini 3が近日中に登場するという噂があり（常に保留中のDeepSeek V4と同様）、業界の期待は現在高まっている。中国のオープンソースモデルの急速な進化により、主要なAI企業は継続的なイノベーションを迫られている。

9. 結論と総合評価

9.1 主要な達成事項

Kimi K2 Thinkingは以下の点で画期的なマイルストーンを達成した:

- 性能の突破:** オープンソースモデルとして初めて、エージェント推論タスクでGPT-5やClaude Sonnet 4.5といったプロプライエタリモデルを上回った
- コスト効率:** 約460万ドルのトレーニングコストで最先端に近い性能を実現し、AI開発の経済性を根本的に変革
- 長期エージェント機能:** 200~300回の連続ツール呼び出しを安定して実行できる能力は、実用的なAIエージェントアプリケーションへの道を開く
- 技術的イノベーション:** INT4量子化とMuonオプティマイザーの成功的な実装により、推論速度とトレーニング効率を大幅に向上

9.2 AI業界への長期的影響

Kimi K2 Thinkingの登場は、オープンソースモデルが単に追いついているだけでなく、複雑なエージェントタスクで性能を直接挑戦している重要な瞬間を示している。既存のリーダーにとって、今後の道のりはベンチマークスコアを超える進化を必要とする。彼らの競争優位性は、生の性能よりも、提供できる信頼性、ユーザーエクスペリエンス、独自の製品統合に依存する可能性がある。

9.3 今後の注目点

- 完全なAgenticモードのリリース:** 近日公開予定の完全機能版が、ベンチマーク性能を実際のユーザーエクスペリエンスでどのように再現するか
- マルチモーダル機能の追加:** テキストのみの制限を克服し、視覚やその他のモダリティへの対応
- グローバル展開:** 現在主に中国市場に焦点を当てているが、国際市場での受容と影響
- 競合の反応:** OpenAI、Anthropic、Googleなどの主要企業がどのように対応するか
- エージェントAI市場の成長:** K2 Thinkingのような能力が実世界のビジネスアプリケーションでどのように活用されるか

9.4 最終評価

Kimi K2 Thinkingは、AI技術の民主化における重要な転換点を表している。オープンソースAI革命において、MiniMax M2とKimi K2 Thinkingは、プロプライエタリの支配の終焉を示している。AI開発の地理的中心が、

オープンソース配布を受け入れる意思のある者に向かってシフトしている証拠である。

この発展は、AI技術へのアクセスがますます民主化される未来を示唆している。6〜12ヶ月遅れていたオープンソースモデルが、初めて重要な推論とエージェントベンチマークでクローズドシステムに匹敵または超える性能を実現した。これは単なる技術的達成ではなく、グローバルAIエコシステムにおける根本的な力関係の変化を示している。

参考資料・情報源

本レポートは、以下の信頼できる情報源に基づいて作成されました:

- Moonshot AI公式Hugging Faceページ
- Moonshot AI公式技術ブログ
- South China Morning Post、CNBC、VentureBeatなどの主要メディア報道
- Interconnects.ai、TechTalksなどのAI業界専門家による分析
- Wikipedia、Crunchbaseなどの企業情報データベース
- 各種技術ベンチマークおよび評価レポート

報告書作成日: 2025年11月12日

最終更新: 2025年11月12日