

2025年11月最新生成AI動向：4大モデル性能徹底比較レポート

作成日: 2025年11月15日 作成者: Manus AI

エグゼクティブサマリー

2025年10月下旬から11月中旬にかけて、生成AIの分野は再び大きな技術的飛躍を遂げました。本レポートでは、この期間に発表された4つの主要な大規模言語モデル（LLM）——**MiniMax-M2**、**Kimi K2 Thinking**、**ERNIE-4.5-VL-28B-A3B-Thinking**、および**GPT-5.1**——について、その性能、アーキテクチャ、革新性を中心に徹底的な調査と分析を行いました。

調査の結果、いくつかの重要なトレンドが明らかになりました。第一に、**Mixture-of-Experts (MoE)** アーキテクチャが業界の標準となり、計算効率とモデル性能の両立を追求する動きが加速しています。第二に、単なる情報生成を超え、複雑なタスクを自律的に解決するための**高度な推論能力**（「**Thinking**」機能）が、各社の差別化要因となっています。第三に、**Baidu**の**ERNIE-4.5-VL**が示すように、マルチモーダル能力は「認識」から「**視覚的推論**」へと深化しています。第四に、**Kimi K2**や**ERNIE-4.5-VL**のような高性能モデルがオープンソースとして公開され、技術革新のエコシステムが大きく変化しつつあります。

各モデルは独自の位置を確立しています。**GPT-5.1**は「アダプティブ・リーズニング」により業界最高水準の性能を維持し、**Kimi K2 Thinking**は驚異的な連続推論能力でエージェントタスクの新たな地平を切り開きました。**ERNIE-4.5-VL**は、軽量ながら高度なマルチモーダル推論を実現し、**MiniMax-M2**は卓越したコスト効率で市場にインパクトを与えています。本レポートが、2025年11月時点における生成AIの最先端技術動向と各モデルの戦略的立ち位置を理解するための一助となれば幸いです。

1. はじめに

2025 年秋、生成 AI の世界はわずか数週間のうちに立て続けに発表された革新的なモデルによって、新たな時代に突入しました。中国の AI スタートアップから世界的なテクノロジー企業まで、各社が次世代の性能を誇るモデルを市場に投入し、その技術競争は激化の一途をたどっています。

本レポートは、2025 年 10 月 26 日から 11 月 12 日にかけてリリースされた以下の 4 つのモデルに焦点を当て、その技術仕様、性能ベンチマーク、そして AI の未来に与える影響を多角的に分析・比較することを目的とします。

- **MiniMax-M2:** 驚異的なコスト効率を誇るエージェント指向モデル
- **Kimi K2 Thinking:** 思考エージェントとして自律的な長期推論を実現したオープンソースモデル
- **ERNIE-4.5-VL-28B-A3B-Thinking:** 「画像で考える」能力を持つ革新的なマルチモーダルモデル
- **GPT-5.1:** 業界のリーダーが示す、推論効率とユーザー体験の新たな最適解

これらのモデルは、それぞれが異なるアプローチで「知能」の限界に挑んでおり、その比較分析を通じて、現在の技術トレンドと今後の発展の方向性を明らかにします。

2. 各モデルの詳細分析

2.1. MiniMax-M2

MiniMax 社が 2025 年 10 月 27 日に発表した **MiniMax-M2** は、特にコスト効率とエージェント能力に焦点を当てて設計されたオープンソースモデルです [1, 2]。総パラメータ数 2300 億 (230B) に対し、アクティブパラメータを 100 億 (10B) に抑えた MoE アーキテクチャを採用しており、軽量でありながら高い性能を発揮します [1]。

性能面では、特にコーディングやエージェント関連のベンチマークで優れた結果を示しています。例えば、実世界のソフトウェア開発タスクを評価する SWE-bench で 69.4、エージェントのツール使用能力を測る τ^2 -Bench で 77.2 という高いスコアを記

録しました [2]。公式発表によると、推論速度は Claude Sonnet 4.5 の約 2 倍でありながら、API 価格はその 8% に過ぎず、圧倒的なコストパフォーマンスを実現しています [3, 4]。主にテキストベースのモデルですが、一部では軽量の視覚タスクもサポートすることが示唆されています。

総評: MiniMax-M2 は、性能とコストのバランスを極限まで追求したモデルであり、特に実用的なアプリケーション開発において、費用対効果を重視する企業にとって非常に魅力的な選択肢となります。

2.2. Kimi K2 Thinking

Moonshot AI 社が 2025 年 11 月 6 日に発表した **Kimi K2 Thinking** は、自律的な長期推論能力を特徴とする画期的なオープンソースモデルです [10, 11]。総パラメータ数 1 兆、活性化パラメータ数 320 億という大規模な MoE アーキテクチャを基盤とし、「思考エージェント」として設計されています [10]。

本モデル最大の特徴は、人間の介入なしに **200~300 ステップに及ぶ連続したツール呼び出し**を安定して実行できる点です [11]。これは、推論の深さを動的にスケールリングする「テスト時スケールリング」技術によって実現されており、複雑な問題解決やエージェントタスクにおいて圧倒的な能力を発揮します。ベンチマークでは、ツール使用を伴う Humanity's Last Exam で 44.9%、エージェント的コーディングを評価する SWE-Bench (複数試行) で 71.6% という SOTA (State-of-the-Art) 級のスコアを達成しました [10]。テキスト専用モデルでありながら、256K トークンという広大なコンテキストウィンドウを持ち、長文理解能力にも優れています。

総評: Kimi K2 Thinking は、自律型 AI エージェントの実現に向けた大きな一歩を示すモデルです。その卓越した長期推論能力は、科学研究、複雑なデータ分析、自律的なソフトウェア開発など、これまで自動化が困難だった領域への応用を可能にします。

2.3. ERNIE-4.5-VL-28B-A3B-Thinking

Baidu が 2025 年 11 月 11 日に発表した **ERNIE-4.5-VL-28B-A3B-Thinking** は、マルチモーダル推論に革命をもたらす可能性を秘めたオープンソースモデルです [16, 17]。

本モデルは、革新的な**「Thinking with Images」**機能を搭載しており、人間のよう
に画像を「見て考える」能力を実現しています [16]。

アーキテクチャの面では、総パラメータ数 300 億に対し、独自の A3B ルーティング
スキームによってアクティブパラメータをわずか 30 億に抑えた効率的な MoE 構造
を採用しています [18]。これにより、軽量の計算プロファイルでありながら、大規
模モデルに匹敵する性能を発揮します。その核となる機能は、画像内の関心領域に
自律的にズームインしたり、内部知識が不足している場合に画像検索ツールを呼び
出したりして、視覚情報を深く理解する能力です [16, 18]。画像だけでなく動画やテ
キストもサポートしており、特にチャートの読解や STEM 分野の視覚的問題解決に
強みを持っています。

*総評: ERNIE-4.5-VL は、マルチモーダル AI が単なる「認識」から高度な「推論」
へと進化する転換点を示しています。その効率性と革新的な視覚的思考能力は、
医療画像診断、科学的発見、インタラクティブな教育など、多岐にわたる分野で
の応用が期待されます。*

2.4. GPT-5.1

OpenAI が 2025 年 11 月 12 日に発表した **GPT-5.1** は、業界最高水準の性能を誇る
GPT-5 のマイナーアップデート版であり、**推論効率とユーザー体験**の向上に焦点を
当てています [24, 25]。本モデルは、クエリの複雑に応じて計算リソースを動的に割
り当てる**「アダプティブ・リーズニング」**という新機能を導入しました [24]。

GPT-5.1 には 2 つのバリエーションが存在します。低レイテンシに最適化された **GPT-5.1
Instant** と、複雑なタスク向けに深い推論を行う **GPT-5.1 Thinking** です [24]。アダプ
ティブ・リーズニングにより、Instant モデルでも中程度の複雑な問題に対応可能に
なり、Thinking モデルは単純なタスクでは高速に応答し、困難なタスクにはより多
くの計算を費やすことで効率化を図っています [26]。また、GPT-5 で課題とされた
命令遵守能力が劇的に改善されており、指定されたフォーマットやスタイルに忠実
に応答する能力が高まっています [28]。コーディングベンチマーク (SWE-bench)
では依然として競合を圧倒しており、総合的な性能で業界のリーダーシップを維持
しています [24]。

総評: GPT-5.1 は、性能の頂点を追求するだけでなく、その知能をいかに効率的かつ快適にユーザーに提供するかに重点を置いた、成熟したモデルです。アダプティブ・リーズニングは、今後の LLM における推論効率の標準となる可能性があります。

3. 主要性能・仕様の横断比較

以下に、4つのモデルの主要な性能と仕様を比較する表を示します。これにより、各モデルの強みと弱み、そして技術的な立ち位置が一目でわかります。

比較項目	MiniMax-M2	Kimi K2 Thinking	ERNIE-4.5-VL	GPT-5.1
開発元	MiniMax	Moonshot AI	Baidu	OpenAI
リリース日	2025年10月27日	2025年11月6日	2025年11月11日	2025年11月12日
ライセンス	オープンソース	オープンソース (Apache 2.0)	オープンソース (Apache 2.0)	プロプライエタリ
アーキテクチャ	MoE (230B/10B)	MoE (1T/32B)	MoE (30B/3B)	Transformer (推定 1-2T)
推論能力	Agentic Workflows	思考エージェント (200-300 ステップ)	Thinking with Images	アダプティブ・リーズニング
マルチモーダル	テキスト中心	テキストのみ	画像・動画・テキスト	テキスト・画像
コンテキスト長	204K トークン	256K トークン	131K トークン	400K トークン
SWE-bench	69.4%	71.6%	-	~74.9%
API コスト (出力)	\$1.20 / 1M トークン	\$0.60 / 1M トークン	(オープンソース)	\$10.00 / 1M トークン
得意分野	コスト効率、コーディング	エージェント的推論、複雑な問題解決	マルチモーダル推論、効率性	総合性能、命令遵守

4. 2025年11月時点の主要技術動向と結論

今回の4大モデルの登場は、2025年後半における生成AIの技術動向を明確に示しています。

- 1 推論能力の深化と多様化:** 全てのモデルが独自の「Thinking」機能を実装し、単なる応答生成から、自律的な問題解決へと進化しています。Kimi K2の「長期推論」、ERNIE-4.5-VLの「視覚的推論」、GPT-5.1の「適応的推論」など、そのアプローチは多様化しており、今後のAIの能力を定義する上での中心的な競争領域となっています。
- 2 効率性の追求 (MoEの標準化):** 大規模なモデル性能を維持しつつ、推論コストをいかに下げるかが重要な課題となっています。アクティブパラメータを大幅に削減するMoEアーキテクチャの採用は、この課題に対する現在の最適解として広く受け入れられています。
- 3 オープンソースの躍進:** Kimi K2やERNIE-4.5-VLのような最先端の性能を持つモデルがオープンソースとして公開されたことは、AIエコシステム全体にとって大きな意味を持ちます。これにより、世界中の開発者が最新技術にアクセスし、新たなイノベーションを生み出す土壌が育まれています。
- 4 マルチモーダルの次なる段階:** ERNIE-4.5-VLが示した「Thinking with Images」は、マルチモーダルAIが、画像や動画の内容を「理解」するだけでなく、それらを基に「推論」し、新たな洞察を生み出す段階に入ったことを示唆しています。

結論として、2025年11月時点の生成AI業界は、性能の絶対的な高さだけでなく、**推論の質、計算効率、そして特定タスクへの専門性**といった多角的な軸で競争が繰り広げられる、より成熟した段階に入ったと言えます。GPT-5.1が依然として総合力でリードする一方、Kimi K2、ERNIE-4.5-VL、MiniMax-M2は、それぞれが独自の強みを武器に、特定領域でリーダーシップを発揮しています。今後の技術開発は、これらの異なるアプローチがどのように融合し、さらに高度な知能へと発展していくのかが焦点となるでしょう。

5. 参考文献

- [1] MiniMax. (2025). *MiniMax-M2*. minimax.io.
- [2] Hugging Face. (2025).*MiniMaxAI/MiniMax-M2*. huggingface.co.
- [3] Microsoft Azure AI Foundry Blog. (2025).*MiniMax M2: The Open-Source Innovator in Coding and Agentic Workflows, Now in Azure*.
- [4] CometAPI. (2025).*MiniMax M2: Why is it the King of Cost-Effectiveness for LLM?*.
- [5] MiniMax Open Platform. (2025).*Pricing*. platform.minimax.io.
- [10] MoonshotAI. (2025).*Kimi-K2*. GitHub.
- [11] Moonshot AI. (2025).*Kimi K2 Thinking*. moonshotai.github.io.
- [12] Moonshot AI Open Platform. (2025).*Pricing*. platform.moonshot.ai.
- [13] ZDNet Japan. (2025).*Moonshot AI、思考エージェント「Kimi K2 Thinking」を発表*.
- [14] Interconnects. (2025).*Kimi K2 Thinking: What It Means*.
- [16] Baidu ERNIE. (2025).*ERNIE-4.5-VL-28B-A3B-Thinking: A Breakthrough in Multimodal AI*.
- [17] Hugging Face. (2025).*baidu/ERNIE-4.5-VL-28B-A3B-Thinking*.
- [18] MarkTechPost. (2025).*Baidu Releases ERNIE-4.5-VL-28B-A3B-Thinking...*
- [19] Baidu Yiyao. (2025).*ERNIE 4.5 Technical Report*.
- [20] Skywork AI. (2025).*Baidu: ERNIE 4.5 VL 28B A3B Free Chat Online*.
- [21] Reddit. (2025).*baidu/ERNIE-4.5-VL-28B-A3B-Thinking released...*
- [22] Galaxy.ai Blog. (2025).*ERNIE 4.5 VL 28B A3B vs. Qwen2.5 7B Instruct*.
- [23] Analytics Vidhya. (2025).*ERNIE-4.5-VL Review: "Thinking with Images" in a 3B Model*.
- [24] Axis Intelligence. (2025).*ChatGPT 5.1 Technical Analysis & Benchmarks*.
- [25] OpenAI. (2025).*GPT-5.1*. openai.com.
- [26] OpenAI. (2025).*GPT-5.1 for developers*.
- [27] OpenAI API. (2025).*Pricing*. openai.com.
- [28] CometAPI. (2025).*What is GPT-5.1 and what updates did it bring?*.
- [29] OpenAI Platform. (2025).*Latest models*.
- [30] AIMLAPI. (2025).*GPT-5.1 Codex*.