

2025年末の最新生成AIモデル調査レポート (MiniMax-M2・Kimi K2 Thinking・ ERNIE-4.5-VL-28B-A3B-Thinking・GPT-5.1)

序論

2025年はオープンソースとフロンティアモデルの競争が激化し、Mixture-of-Experts (MoE) や長コンテキスト対応の大型モデルが次々に登場した。本レポートでは、2025年10月下旬～11月に相次いでリリースされた **MiniMax-M2**、**Kimi K2 Thinking**、**ERNIE-4.5-VL-28B-A3B-Thinking**、**GPT-5.1** という4つの最新生成AIモデルについて、公式情報・技術ニュースを用いて性能を徹底的に調査・比較する。

MiniMax-M2 (2025年10月26日リリース)

主要仕様と特徴

- **アーキテクチャ** – MiniMax-M2 はミクスチャー-オブ-エキスパート (MoE) 方式を採用。総パラメータ数は **200 B** で、推論時にアクティブになるのは **10 B** のみ ¹。パラメータ効率のおかげで **4枚の H100 GPU** で動作し、前世代 (M1) より推論速度を約2倍向上 ²。
- **エージェント適性** – 公式発表では、M2 はエージェント用のテキストモデルとして設計され、ツール利用や長い計画の実行に優れると述べている ³。VentureBeat の記事は、M2 が多くのエージェント系ベンチマークで高いスコア (τ²-Bench 77.2、BrowseComp 44.0、FinSearchComp 65.5 など) を出し、「多くの独自モデルに近い性能」を示したと報告 ⁴。
- **オープンソースとコスト** – モデルは MIT ライセンス下で公開されており、Hugging Face や GitHub から入手可能。MiniMax 社の発表によれば、1 M入力トークン当たり **0.30 米ドル**、出力トークン当たり **1.20 米ドル** と、Claude Sonnet 4.5の約8%の価格で提供される ³。推論速度は **約100 TPS** とされる ³。
- **コンテキストウィンドウ** – 比較サイト Galaxy.ai によると、MiniMax-M2 は **約204.8k トークンの入力コンテキスト** と **131.1k トークンの出力** に対応し ⁵、長文処理に強い。

性能評価と改善点

M2 は t²-Bench や BrowseComp のようなエージェントベンチマークで優秀なスコアを示し、同社は「工具利用と深い検索能力は海外トップモデルに迫り、プログラミングでは若干劣るものの国内モデルではトップクラス」と説明している ³。VentureBeat の評価では、SWE-bench Verified 69.4%、ArtifactsBench 66.8% など、コード修正やエージェント環境でも高い成績を残している ⁴ ⁶。

総じて、MiniMax-M2 は **大規模MoEでコスト効率の良いエージェント向けモデル**として位置付けられる。前世代比では推論速度とコストが大きく改善し、長コンテキストと計画能力が強化された。

Kimi K2 Thinking (2025年11月6日リリース)

主要仕様と特徴

- **アーキテクチャ** - Kimi K2 Thinking は Moonshot AI による reasoning 用オープンソースモデルで、**1 T (約1兆) パラメータ**の MoE だがアクティブパラメータは **32 B** に抑えられている ⁷。61層・7168次元の隠れ層・384人のエキスパートを持ち、**256 kトークンのコンテキストウィンドウ**に対応 ⁸。
- **思考プロセス (Thinking 機能)** - モデルは計画→行動→検証→反省→改善という **インタリーブ・リゾニング**を組み込み、**200~300回の連続的なツール呼び出し**を安定して実行する ⁹。
Heavy Modeでは8本の推論パスを並列探索し、難問で精度を高める ¹⁰。
- **量子化と速度** - K2 Thinking は **INT4量子化**が標準となっており、Weight Only QAT によって標準モデルより **約2倍高速**に推論できる ¹¹。
- **透明な推論** - API では reasoning トレースを返し、ユーザーが思考の各ステップを確認できる。DataCamp によると、HLE (w/ tools) 44.9 %、BrowseComp 60.2 %、SWE-bench Verified 71.3 % などのベンチマークで GPT-5 (high) を上回る ⁷ ¹²。
- **価格と利用形態** - DataCamp が紹介する標準プランでは入力トークン **0.60 USD/M**、出力トークン **2.50 USD/M**、推論速度は **約18 TPS** で、**ターボプラン**では **1.15 USD/M 入力・8.00 USD/M 出力、85 TPS** の高速化が可能 ¹³。

性能評価と特徴比較

K2 Thinking は人類最後の試験 HLE や数理コンテスト AIME25 で GPT-5 high を上回り、BrowseComp や FinSearchComp-T3 でも優れた結果を示した ⁷ ¹²。さらに **LiveCodeBench** と **SWE-bench Verified** で 71.3 % の正解率を達成し、オープンモデルとしてトップクラスのコーディング能力を示した ¹⁴。長期的なタスクやエージェント型のワークフローに強く、標準プランでは他社より低価格で利用できる点も魅力である。

ERNIE-4.5-VL-28B-A3B-Thinking (2025年11月11日リリース)

主要仕様と特徴

- **アーキテクチャ** - 百度が公開したマルチモーダルモデルで、**28 Bパラメータ** (VL ブランチ) のうち **3 B** だけを推論時に有効化する **A3B ルーティング**が特徴 ¹⁵。これにより **30 B 級モデル並みの能力を 3 B モデルの計算コスト**で実現し、単一の **80 GB GPU**でも動作する ¹⁶。
- **学習と強化学習** - 大規模な視覚言語推論データで中間学習を行い、**GSPO** と **IcePop** を使った **マルチモーダル強化学習**および難易度動的サンプリングにより表現と意味整合を強化した ¹⁷ ¹⁵。
- **Thinking with Images** - 人間のように画像を自由にズームイン/アウトし細部を掴む「**Thinking with Images**」機能を備える ¹⁸。画像検索など外部ツールとの連携により長尾知識を取得し、視覚的推論やチャート解析などに強い ¹⁸。
- **推論能力** - モデルカードは、視覚推論、STEM 課題、視覚的グラウンディング、画像との思考、ツール利用、動画理解など6つの能力を列挙し、トップモデルに匹敵する性能を公表している ¹⁹。MarkTechPost は ERNIE が Qwen-2.5-VL 7B/32B と同等以上の性能を示したと報じた ²⁰。
- **コンテキストウィンドウと価格** - Galaxy.ai の比較ページによると、**入力約131.1k トークン、出力約65.5k トークン**に対応し、価格は **入力0.07 USD/M、出力0.28 USD/M**と非常に低価格 ⁵ ²¹。Apache 2.0 ライセンスで公開され、Hugging Face などから利用できる。

特色と位置付け

ERNIE-4.5-VL-28B-A3B-Thinking は、**視覚と言語を統合したエージェント向けモデル**として突出している。3 B アクティブパラメータで大幅な推論効率を実現しながら、視覚推論・STEM 問題解決・動画理解などが可

能であり、他のテキスト主体モデルでは困難なシナリオをカバーする。百度は他社の Gemini 2.5 Pro や GPT-5 シリーズを上回る性能を主張しているが独立検証は限定的である ²²。

GPT-5.1 (2025年11月12日リリース)

主要仕様と特徴

- **バージョン構成** – OpenAI は ChatGPT 向けに **GPT-5.1 Instant** (会話重視) と **GPT-5.1 Thinking** (高度な推論) の2モデルを提供。開発者向け API には reasoning モデルとして `gpt-5.1` と `gpt-5.1-chat` が追加された。
- **適応的推論** – GPT-5.1 は **adaptive reasoning** でタスク難易度に応じて推論量を調整し、単純なタスクでは GPT-5 より **2~3倍高速**、複雑なタスクでも同等の性能を維持する ²³。
- **新ツール** – 開発者向けに `apply_patch` (自由形式のコード修正) と `shell` (ローカルコマンド実行) ツールを追加 ²⁴。
- **コンテキストウィンドウ** – Microsoft Azure のモデル一覧によると、`gpt-5.1` は **約400kトークン** (入力272k + 出力128k) まで扱える ²⁵。OpenRouter も 400k トークンと記載し、API 価格は **入力 1.25 USD/M、出力 10 USD/M** としている ²⁶。
- **キャッシュの拡張** – 従来数分だったプロンプトキャッシュが **24時間保持** できるようになり、長時間の対話や長期タスクでコストと遅延を削減 ²⁷。

性能評価

OpenAI の開発者ブログは GPT-5.1 と GPT-5 (high) の比較を公表している。SWE-bench Verified では **76.3%** と GPT-5 の 72.8% を上回り、GPQA Diamond で **88.1%**、MMMU で **85.4%** など多くのベンチマークで改善が確認された ²⁸。AIME 2025 や Tau2-bench Retail では GPT-5 が僅差で上回るが、総合的には GPT-5.1 が優勢である。また開発者企業からは「**2~3倍高速になり、半分のトークンで同等以上の結果**」との声が報告されている ²³。

四モデルの比較

コンテキストウィンドウ・費用・アーキテクチャ比較

モデル	アクティ ブパラ メータ /総パ ラメー タ	コンテキ ストウ ィンドウ (入力/出 力)	API コスト(入力/出 力)	備考
MiniMax-M2	10 B / 200 B ¹	約 204.8k / 131.1k トークン ⁵	\$0.30 / \$1.20 per M tokens ³	エージェント向 けテキスト LLM。H100×4 で動作。MITラ イセンス。
Kimi K2 Thinking	32 B / 1 T ⁷	256k / 約 256k ⁸	\$0.60 / \$2.50 (M標 準) ¹³ 、ターボは \$1.15 / \$8.00	200-300回の ツール呼び出 し・INT4量子 化・思考過程出 力 ⁹ 。

モデル	アク ティブ パラ メータ ／総パ ラメー タ	コンテキ ストウィ ンドウ (入力/出 力)	API コスト(入力/出 力)	備考
ERNIE-4.5-VL-28B-A3B-Thinking	3 B / 28 B <small>15</small>	131.1k / 65.5k <small>5</small>	\$0.07 / \$0.28 per M <small>21</small>	マルチモーダル (画像・動画・ テキスト)。 Apache 2.0。
GPT-5.1	非公表 (推定 数百 B) / -	272k / 128k (合 計400k) <small>25</small>	\$1.25 / \$10.00 per M <small>26</small>	アダプティブ推 論・prompt cache 24h・ apply_patch と shell ツール <small>29</small> 。

得意分野・評価

項目	MiniMax-M2	Kimi K2 Thinking	ERNIE-4.5-VL-28B-A3B-Thinking	GPT-5.1
コーディング	SWE-bench Verified 69.4 %、ArtifactsBench 66.8 %、GAIA 75.7 <small>4</small> 。プログラミング性能はトップモデルに僅かに劣るがオープンソースでは上位。	SWE-bench Verified 71.3 %、LiveCodeBench 74 %前後 <small>14</small> 。プログラミングタスクでGPT-5 (high) に近い。	コーディングベンチは公開されていないが、視覚情報の解析による STEM 問題解決に強い <small>19</small> 。	SWE-bench Verified 76.3 % <small>28</small> 。新しい <code>apply_patch</code> ツールや adaptive reasoning によりコード修正が迅速。
論理推論・数学	τ^2 -Bench 77.2・BrowseComp 44.0 <small>4</small> 。長い計画やツール利用を伴うエージェントタスクで高評価。	HLE (with tools) 44.9 %、AIME25 84.3、BrowseComp 60.2 <small>7 12</small> 。計画・反省機能によって多段推論が得意。	STEM 問題や図表解析を視覚情報と統合して解く機能 <small>18</small> 。	GPQA Diamond 88.1 %、AIME 2025 94.0%、Tau2-bench Airline 67.0 <small>28</small> 。多数の基準で GPT-5 より向上。
創作・自然言語	テキストのみだが、長文コンテンツで小説や記事を生成可能。	日本語・中国語を含む多言語で自然な文章を生成。思考過程を表示できる。	画像と文章を組み合わせた物語や説明が得意。動画理解やチャート解析も可能 <small>19</small> 。	Instant モードは会話・創作に強く、Thinking モードは論理・分析に強い。トーンの制御やカスタマイズが可能。

項目	MiniMax-M2	Kimi K2 Thinking	ERNIE-4.5-VL-28B-A3B-Thinking	GPT-5.1
エージェント & ツール呼び出し	外部API呼び出し・長期計画に優れ、BrowseComp や FinSearchComp で高スコア ⁴ 。	200-300連続ツール呼び出しが可能で、検証・反省ステップを含むインタリーブ推論 ⁹ 。	画像検索などツール統合に対応し、視覚的な長尾知識取得に強い ¹⁸ 。	<code>apply_patch</code> や <code>shell</code> など新ツールを提供し、コード編集やローカルコマンド実行が容易 ²⁴ 。
長文処理・速度	204k入力で長い文書を扱える。推論速度は約100 TPS ³ 。	256kコンテキスト。INT4量子化により推論速度が標準の2倍 ¹¹ 。	131k入力でも視覚処理を含む。A3Bルーティングにより3Bアクティブで効率的 ¹⁵ 。	400kコンテキスト。adaptive reasoningにより簡単なタスクは従来より2~3倍高速 ²³ 。
価格	入力0.30 USD/M、出力1.20 USD/M ³ 。	標準0.60/2.50、ターボ1.15/8.00 USD/M ¹³ 。	0.07/0.28 USD/M ²¹ 。	1.25/10 USD/M ²⁶ 。プロンプトキャッシュ利用で安価になるケースも ²⁷ 。

総合考察と技術動向（2025年11月時点）

- MoEの主流化と効率重視** - 4モデルすべてが Mixture-of-Experts を採用し、総パラメータは数百億~1兆規模だが、推論時には一部のみを活性化。これにより、高性能を保ちつつ、コストとレイテンシを大幅に削減している。特に ERNIE-4.5 は 3 B アクティブで 80 GB GPU 上で動作し、MiniMax-M2 も 10 B アクティブで 4 × H100 で済むなど、インフラ要件が緩和された。
- 長コンテキスト競争** - GPT-5.1 の 400k トークンや K2 Thinking の 256k など、2025年にはコンテキストウィンドウの拡大が顕著となり、長いドキュメントや多ターンの会話を単一セッションで処理できるようになった。一方、ERNIE-4.5 は視覚情報を含む 131k 入力で実用性を保ちつつ、推論コストを低く抑えている。
- エージェント化とツール統合** - MiniMax-M2 と K2 Thinking はツール呼び出し数や連続計画に重点を置き、実務に近いタスクを自律的に処理する。GPT-5.1 も `apply_patch` や `shell` を導入し、エージェント型モデルとしての能力を強化。ERNIE-4.5 は画像検索と連携して長尾知識を補完し、マルチモーダルエージェントの先駆けとなっている。
- コーディングおよび数学性能の向上** - 4モデル共にコード修正や数学問題で高得点を記録したが、GPT-5.1 が SWE-bench Verified 76.3% とトップを維持し、K2 Thinking がそれに続く。MiniMax-M2 もオープンモデルとして優秀で、ERNIE-4.5 は視覚情報を使った STEM 問題解決で差別化している。
- 価格差とライセンス** - コスト面では ERNIE-4.5 と MiniMax-M2 が非常に低価格で提供され、オープンソース (MIT/Apache 2.0) であるのに対し、GPT-5.1 は従量課金が高めだが高性能と柔軟性を提供する。K2 Thinking は標準とターボの2プランを用意し、使い分けによってコスト効率を調整できる。

結論

2025年11月時点の生成AI市場では、**MiniMax-M2** が低コストかつ高いエージェント性能を持つオープンソースモデルとして存在感を示し、**Kimi K2 Thinking** がツール利用を前提とした長期推論や複雑な論理タスクでフロンティアモデルに迫る実力を見せた。**ERNIE-4.5-VL-28B-A3B-Thinking** は視覚と言語の統合によって画像・動画解析を可能にし、低コストで企業向けに魅力的な選択肢となっている。一方、**GPT-5.1** は開発者向

けの高度な推論モデルとして最先端の総合性能と柔軟性を提供し、長コンテキスト処理とエージェントツール統合を強化した。

今後は、MoE アーキテクチャの効率化に加え、視覚・音声など複数モダリティを扱うモデルの進化、そして推論過程の透明性や自律エージェントの安全な実装が主な研究課題となるだろう。

1 MiniMax M2 Benchmarks & Analysis

<https://artificialanalysis.ai/articles/minimax-m2-benchmarks-and-analysis>

2 3 MiniMax M2

<https://www.minimax.io/news/minimax-m2>

4 6 MiniMax-M2 is the new king of open source LLMs (especially for agentic tool calling) | VentureBeat

<https://venturebeat.com/ai/minimax-m2-is-the-new-king-of-open-source-llms-especially-for-agentic-tool>

5 21 ERNIE 4.5 21B A3B Thinking vs MiniMax M2 (Comparative Analysis) | Galaxy.ai

<https://blog.galaxy.ai/compare/ernie-4-5-21b-a3b-thinking-vs-minimax-m2>

7 10 12 Kimi K2 Thinking: Open-Source LLM Guide, Benchmarks, and Tools | DataCamp

<https://www.datacamp.com/tutorial/kimi-k2-thinking-guide>

8 14 moonshotai/Kimi-K2-Thinking · Hugging Face

<https://huggingface.co/moonshotai/Kimi-K2-Thinking>

9 Introducing Kimi K2 Thinking, China's 'Most Capable' Open-Source Model

<https://www.vktr.com/ai-market/introducing-kimi-k2-thinking-chinas-most-capable-open-source-model/>

11 5 Thoughts on Kimi K2 Thinking - by Nathan Lambert

<https://www.interconnects.ai/p/kimi-k2-thinking-what-it-means>

13 The biggest model update this week wasn't GPT-5.1, it was Kimi K2: AI Update #3

<https://www.news.aakashg.com/p/the-biggest-model-update-this-week>

15 20 Baidu Releases ERNIE-4.5-VL-28B-A3B-Thinking: An Open-Source and Compact Multimodal Reasoning Model Under the ERNIE-4.5 Family - MarkTechPost

<https://www.marktechpost.com/2025/11/11/baidu-releases-ernie-4-5-vl-28b-a3b-thinking-an-open-source-and-compact-multimodal-reasoning-model-under-the-ernie-4-5-family/>

16 18 22 Baidu just dropped an open-source multimodal AI that it claims beats GPT-5 and Gemini | VentureBeat

<https://venturebeat.com/ai/baidu-just-dropped-an-open-source-multimodal-ai-that-it-claims-beats-gpt-5>

17 19 baidu/ERNIE-4.5-VL-28B-A3B-Thinking · Hugging Face

<https://huggingface.co/baidu/ERNIE-4.5-VL-28B-A3B-Thinking>

23 24 27 28 29 Introducing GPT-5.1 for developers | OpenAI

<https://openai.com/index/gpt-5-1-for-developers/>

25 Azure OpenAI reasoning models - GPT-5 series, o3-mini, o1, o1-mini - Azure OpenAI | Microsoft Learn

<https://learn.microsoft.com/en-us/azure/ai-foundry/openai/how-to/reasoning>

26 GPT-5.1 - API, Providers, Stats | OpenRouter

<https://openrouter.ai/openai/gpt-5.1>