

# 2025年11月最新生成AIモデル性能比較レポート

調査日：2025年11月15日

## 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、AI業界は驚異的な進化を遂げた。中国発のオープンソースモデル（MiniMax-M2、Kimi K2 Thinking、ERNIE-4.5-VL-28B-A3B-Thinking）とOpenAIのGPT-5.1が相次いでリリースされ、性能面でも価格面でもパラダイムシフトが起きている。

### 主要な発見

- オープンソースの台頭:** 中国企業のオープンソースモデルが、プロプライエタリモデル（GPT-5、Claude Sonnet 4.5）と同等以上のベンチマーク性能を達成
- コスト効率の革命:** MiniMax-M2はClaude Sonnetの8%のコストで競合性能を実現
- 推論モデルの成熟:** "Thinking"機能を持つモデルが主流化し、複雑なタスクでの性能が飛躍的に向上
- マルチモーダルな進化:** ERNIE-4.5-VLは3B活性パラメータで大規模モデルに匹敵する画像理解能力を実現

## 1. MiniMax-M2（2025年10月27日リリース）

### 1.1 基本仕様

開発企業: MiniMax（中国・上海）

リリース日: 2025年10月27日

アーキテクチャ: Mixture-of-Experts (MoE)

パラメータ数:

- 総パラメータ: 230B（2,300億）
- 活性パラメータ: 10B（100億/トークン） コンテキストウィンドウ: 204,800トークン（約200K）
- ライセンス: MIT（オープンソース）

### 1.2 技術的特徴

MiniMax-M2は、エージェントワークフローとコーディングタスクに特化した設計が特徴。MoEアーキテクチャにより、推論時には10Bパラメータのみを活性化することで、高い効率性を実現している。

### 主要な技術革新:

- インターリーブドThinking:** `<think>...</think>` ブロックで内部推論プロセスを明示化
- マルチツール統合:** MCP（Model Context Protocol）、シェル、ブラウザ、Pythonインタプリタをネイティブサポート

- **高速推論:** 約100トークン/秒（Claude Sonnet 4.5の約2倍）

1.3 ベンチマーク性能

ベンチマーク	MiniMax-M2	GPT-5	Claude 4.5	備考
SWE-bench Verified	69.4%	74.9%	77.2%	コード修正タスク
Terminal-Bench	46.3%	43.8%	50.0%	ターミナル操作
BrowseComp	44.0%	54.9%	19.6%	ブラウザエージェント
Multi-SWE-Bench	36.2%	-	-	マルチファイル編集
$\tau^2$ -Bench	77.2%	-	-	エージェントツール使用
Intelligence Index	61	68	-	Artificial Analysis総合指標

1.4 価格・可用性

API価格:

- 入力: \$0.30/100万トークン
- 出力: \$1.20/100万トークン
- **Claude Sonnet 4.5と比較して約8%のコスト**

デプロイメント:

- Hugging Faceで完全なモデルウェイトを公開
- vLLM、SGLangによるローカル実行をサポート
- FP32、BF16、FP8形式で提供

1.5 強みと弱み

強み:

- コストパフォーマンスが極めて高い
- エージェントワークフローでの優れたパフォーマンス
- オープンソースで完全にカスタマイズ可能
- Claudeを上回るブラウザエージェント能力

弱み:

- テキストのみ対応（マルチモーダル非対応）
  - 非常に冗長（120M トークン使用でIntelligence Index評価を完了、他モデルより多い）
  - コーディングベンチマークではGPT-5やClaudeにやや劣る
-

## 2. Kimi K2 Thinking（2025年11月6日リリース）

### 2.1 基本仕様

開発企業: Moonshot AI（中国・北京）

リリース日: 2025年11月6日

アーキテクチャ: Mixture-of-Experts (MoE)

パラメータ数:

- 総パラメータ: 1T（1兆）
- 活性パラメータ: 32B（320億/トークン） コンテキストウィンドウ: 256,000トークン  
トレーニングコスト: \$4.6M（460万ドル）  
トレーニングデータ: 15.5兆トークン  
ライセンス: Modified MIT（大規模商用利用時に帰属表示必須）

### 2.2 技術的特徴

Kimi K2 Thinkingは、「Thinking Agent」として設計された初のモデルで、推論しながら同時にツールを使用できる点が革新的。

主要な技術革新:

- ネイティブINT4量子化: 量子化を考慮した学習により、INT4推論で約2倍の速度向上
- テストタイム拡張: 推論トークンとツール呼び出し回数を動的に拡張（200-300回の連続ツール呼び出し可能）
- MuonClipオプティマイザー: 1兆パラメータスケールでの勾配安定化
- 3階層メモリ構造: 短期・中期・長期の階層的コンテキスト管理

### 2.3 ベンチマーク性能

ベンチマーク	Kimi K2 Thinking	GPT-5	Claude 4.5	特記事項
HLE (Humanity's Last Exam)	44.9%	41.7%	-	最難関マルチドメイン試験
BrowseComp	60.2%	54.9%	24.1%	K2がトップ
GPQA Diamond	85.7%	84.5%	-	大学院レベル科学
AIME 2025	~94%	~94%	-	数学オリンピック
HMMT 2025	~95%	~94%	-	ハーバード数学大会
SWE-bench Verified	71.3%	74.9%	-	コード修正
LiveCodeBench v6	83.1%	-	-	競技プログラミング
τ <sup>2</sup> -Bench Telecom	93%	-	-	最高スコア

### 2.4 価格・可用性

API価格:

- 入力: \$0.55/100万トークン
- 出力: \$2.25/100万トークン

- GPT-5 (\$1.25/\$10) の約1/10のコスト

#### アクセス方法:

- OpenRouter経由でアクセス可能
- Hugging Faceでモデルウェイト公開 (594GB)
- ローカルデプロイには8×H200 GPU推奨 (約250GB統合メモリ)

## 2.5 強みと弱み

#### 強み:

- **多くのベンチマークでGPT-5を上回る:** 特にエージェント推論とマルチステップ検索で優位
- 透明性の高い推論プロセス (thinking trace を確認可能)
- 圧倒的なコストパフォーマンス
- 競技プログラミングで優れた性能
- オープンウェイトでカスタマイズ可能

#### 弱み:

- リポジトリレベルのエンジニアリングではGPT-5がやや優位
- 一部のタスクでやや冗長なコード生成の傾向
- マルチモーダル非対応 (テキストのみ)

## 2.6 業界への影響

Kimi K2 Thinkingは、オープンソースモデルがプロプライエタリモデルの性能を上回った象徴的な例として注目されている。CNBCの報道によれば、トレーニングコストはわずか\$4.6Mで、これはOpenAIやAnthropicが数億ドルを投資していることと対照的。

---

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

### 3.1 基本仕様

**開発企業:** Baidu (中国・北京)

**リリース日:** 2025年11月11日

**アーキテクチャ:** Heterogeneous Multimodal MoE (異種マルチモーダルMoE)

#### パラメータ数:

- 総パラメータ: 30B (約280億)
  - 活性パラメータ: 3B (30億/トークン) **モダリティ:** テキスト + 画像 (Vision-Language)
- ライセンス:** Apache 2.0 (完全オープンソース)

### 3.2 技術的特徴

ERNIE-4.5-VL-28B-A3B-Thinkingは、Baiduが開発したマルチモーダル推論モデル。特筆すべきは、わずか3Bの活性パラメータで業界トップクラスの性能を実現している点。

主要な技術革新:

- **Thinking with Images:** 推論プロセス中に画像を拡大・縮小して詳細を確認
- **視覚言語推論コーパス:** 大規模な視覚言語推論データで中間訓練
- **GSPO + IcePop戦略:** マルチモーダル強化学習による動的難易度サンプリング
- **視覚グラウンディング:** 産業シナリオでの精密な物体検出と指示実行

3.3 ベンチマーク性能

Baiduの公式発表によれば、ERNIE-4.5-VL-28B-A3B-Thinkingは以下の領域で優れた性能を示す:

能力領域	性能	競合比較
視覚推論	高	Gemini 2.5 Proを上回ると主張
STEM推論	高	画像からの問題解決に優位
視覚グラウンディング	高	産業シナリオで精密
文書理解	高	密集したテキストとチャートに強い
ビデオ理解	対応	マルチフレーム解析可能

具体的な比較（VentureBeatの報道）:

- VQA、MMBench、SEED-BenchでGemini 2.5 Proを上回る
- 7B+のオープンモデルと同等以上の性能

3.4 価格・可用性

推論要件:

- 最小GPU: 80GB VRAM（基本起動）
- wint8量子化使用時: 約60GB VRAM
- 推論速度: 通常モデルの2-3倍（3B活性パラメータのため）

デプロイメント:

- vLLM、FastDeployでサーバー化可能
- ERNIEKitでファインチューニング対応
- Hugging Faceで完全公開

3.5 強みと弱み

強み:

- **超軽量:** 3B活性パラメータで大規模モデルに匹敵
- マルチモーダル対応（他の3モデルはテキストのみ）
- Apache 2.0ライセンスで商用利用自由
- 「Thinking with Images」による視覚的推論の透明性

- 推論速度が速い

弱み:

- 総合的なベンチマーク比較データが限定的
- 第三者による独立した検証が不足
- 一部タスク（指の数え間違いなど）で精度に課題

4. GPT-5.1（2025年11月12日リリース）

4.1 基本仕様

開発企業: OpenAI（米国）  
リリース日: 2025年11月12日  
モデルバリエーション:

- GPT-5.1 Instant: 高速会話型
- GPT-5.1 Thinking: 高度推論型 コンテキストウィンドウ: 最大196K（Thinking、有料プラン） / APIでは400K入出力合計  
ライセンス: プロプライエタリ（APIのみ）

4.2 技術的特徴

GPT-5.1は、GPT-5（2025年8月リリース）からわずか3ヶ月での改良版。アーキテクチャの革新よりも、UXと実用性の向上に焦点を当てている。

主要な改善点:

- 適応的推論（Adaptive Reasoning）: Instantモデルが複雑なクエリを自動検出し、必要に応じて深い推論を実行
- 動的リソース配分: 簡単なタスクで57%トークン削減、複雑なタスクで71%増加
- 8種類のパーソナリティプリセット: Professional、Candid、Quirky、Friendly、Efficient、Nerdy、Cynical、Default
- 指示追従の向上: より正確にユーザーの指示に従う

4.3 ベンチマーク性能

OpenAIは詳細なベンチマーク数値を公開していないが、以下の改善を主張:

ベンチマーク	GPT-5.1	GPT-5	改善点
SWE-bench Verified	76.3%	72.8%	+3.5pt（OpenAI公式）
AIME 2025	「大幅改善」	100%*	*thinking+Python使用時
Codeforces	「顕著な向上」	-	詳細数値非公開

速度性能（GPT-5.1 Thinking vs GPT-5 Thinking）:

- 最も簡単なタスク: 約2倍高速

- 最も複雑なタスク: 約2倍低速（より深く考える）

4.4 価格・可用性

API価格（予想）：

- GPT-5の価格を維持: \$1.25/M入力、\$10/M出力
- GPT-5.1 Thinkingは複雑タスクで71%多くトークンを使用するため、実質コストは上昇の可能性

アクセス:

- ChatGPT Pro、Plus、Go、Business: 即時利用可能
- Freeユーザー: 段階的ロールアウト
- API: `gpt-5.1-chat-latest` (Instant)、`GPT-5.1` (Thinking)

4.5 強みと弱み

強み:

- SWE-bench Verifiedで最高スコア（76.3%）
- 会話の自然さとトーンカスタマイズ
- 既存のOpenAIエコシステムとの完全統合
- 企業向けサポート体制

弱み:

- 価格が高い: 中国モデルの10-20倍のコスト
- ベンチマーク透明性の低下（詳細数値非公開）
- 安全性の後退: システムカードによれば、13のうち9カテゴリで安全性が低下
- オープンソースではない

4.6 市場への影響

GPT-5.1のリリースは、OpenAIが中国のオープンソースモデルからの圧力に対応している証拠と見られている。性能向上は漸進的で、価格優位性を失いつつあることから、「ユーザー体験」を差別化要因として強調している。

5. 横断的比較分析

5.1 性能マトリックス

モデル	コーディング	エージェント推論	数学	効率性	マルチモーダル
MiniMax-M2	★★★★★	★★★★★★	★★★★	★★★★★★	×
Kimi K2 Thinking	★★★★★★	★★★★★★	★★★★★★	★★★★★	×

		エ   ジェント推			

5.2 コストパフォーマンス比較

100万トークン入力 + 100万トークン出力の場合のコスト:

モデル	入力コスト	出力コスト	合計	GPT-5比
MiniMax-M2	\$0.30	\$1.20	\$1.50	13%
Kimi K2 Thinking	\$0.55	\$2.25	\$2.80	25%
ERNIE-4.5-VL	-	-	推定\$1-2	-
GPT-5.1	\$1.25	\$10.00	\$11.25	100%

結論: 中国モデルは、GPT-5.1の1/4~1/8のコストで同等以上の性能を提供。

5.3 使用シナリオ別推奨

コーディング・ソフトウェア開発

推奨: GPT-5.1 > Kimi K2 Thinking > MiniMax-M2

- **GPT-5.1:** SWE-bench最高スコア、リポジトリレベルの理解に優位
- **Kimi K2:** 競技プログラミングと簡潔なコード生成に強い
- **MiniMax-M2:** マルチファイル編集とCI/CDパイプライン統合に適合

エージェントワークフロー・自律タスク

推奨: Kimi K2 Thinking > MiniMax-M2 > GPT-5.1

- **Kimi K2:** BrowseCompで60.2%、200-300回の連続ツール呼び出し可能
- **MiniMax-M2:**  $\tau^2$ -Benchで77.2%、高速なエージェント処理
- **GPT-5.1:** より保守的だが信頼性重視の環境に適合

数学・論理推論

推奨: Kimi K2 Thinking  $\approx$  GPT-5.1 > MiniMax-M2

- **Kimi K2:** AIME、HMMTで94-95%
- **GPT-5.1:** AIME 2025で「大幅改善」を主張
- **MiniMax-M2:** 基本的な数学能力は十分だが特化していない

マルチモーダル（画像+テキスト）

推奨: ERNIE-4.5-VL（唯一の選択肢）



- 他3モデルはテキストのみ対応
- 3B活性パラメータで軽量かつ高速

## コスト重視・大規模デプロイ

推奨: MiniMax-M2 > Kimi K2 Thinking > ERNIE-4.5-VL

- **MiniMax-M2:** 最も安価（\$1.50/2M tokens）
- **Kimi K2:** 性能とコストのバランスが優秀
- **ERNIE-4.5-VL:** 軽量で推論速度が速い

## 5.4 技術トレンドの分析

### オープンソースの優位性拡大

- **2024年末:** オープンソースはGPT-5に18ポイント差で劣る
- **2025年11月:** Kimi K2がGPT-5を複数ベンチマークで上回る
- **傾向:** 性能ギャップがゼロに収束、コストギャップは拡大

### MoE（Mixture-of-Experts）の主流化

全4モデルがMoE構造を採用し、効率性を追求:

- **MiniMax-M2:** 230B → 10B活性
- **Kimi K2:** 1T → 32B活性
- **ERNIE-4.5-VL:** 30B → 3B活性
- **GPT-5.1:** 詳細非公開だが類似構造と推測

### 「Thinking」機能の標準化

推論プロセスを可視化・最適化する「Thinking」機能が主流に:

- **インターリーブドThinking:** MiniMax-M2、Kimi K2
- **適応的推論:** GPT-5.1 Instant
- **Thinking with Images:** ERNIE-4.5-VL

### 中国AI企業の急速な台頭

- DeepSeek R1（2025年1月）がきっかけ
- わずか6ヶ月で複数の企業がフロンティアモデルに到達
- トレーニングコストの劇的削減（MiniMax-M1: \$534K、Kimi K2: \$4.6M）

---

## 6. 実用的な仕様・展開情報

### 6.1 コンテキストウィンドウ比較

モデル	コンテキスト長	実用的な用途
MiniMax-M2	204K	長文ドキュメント分析
Kimi K2 Thinking	256K	企業規模のテキスト分析、研究レポート
ERNIE-4.5-VL	131K（推定）	マルチモーダル長文処理
GPT-5.1	196K (Thinking)	包括的なコンテキスト理解

歴史的比較:

- MiniMax-01（2025年1月）: 4M（400万）トークン - 業界最長記録
- Gemini 1.5 Pro: 2M トークン
- Claude 3 Opus: 200K トークン

6.2 推論速度

モデル	トークン/秒	レイテンシ	備考
MiniMax-M2	~100	低	Claude Sonnetの2倍
Kimi K2 Thinking	-	中	INT4量子化で高速化
ERNIE-4.5-VL	-	低	3B活性で通常の2-3倍
GPT-5.1 Instant	-	低	簡単タスクで2倍高速
GPT-5.1 Thinking	-	高	複雑タスクで2倍低速

6.3 多言語対応



モデル	主要対応言語	日本語性能
MiniMax-M2	英語、中国語中心	限定的
Kimi K2 Thinking	多言語（中国語特化）	良好
ERNIE-4.5-VL	中国語、英語（視覚言語）	良好
GPT-5.1	80+言語	優秀



6.4 デプロイメントオプション

クラウドAPI

- **MiniMax-M2:** MiniMax公式API、OpenRouter
- **Kimi K2:** Moonshot API、OpenRouter
- **ERNIE-4.5-VL:** Baidu API
- **GPT-5.1:** OpenAI API、Azure OpenAI

オンプレミス/ローカル

- **MiniMax-M2:**  vLLM、SGLang（4×H100推奨）
- **Kimi K2:**  vLLM、MLX（8×H200推奨）

- **ERNIE-4.5-VL:**  vLLM、FastDeploy (80GB VRAM)
  - **GPT-5.1:**  APIのみ
- 

## 7. 2025年11月時点での生成AI業界動向

### 7.1 パラダイムシフト

#### 1. オープンソースの性能逆転

- 従来: プロプライエタリ > オープンソース
- 現在: オープンソース (Kimi K2) がGPT-5を上回る

#### 2. 価格競争の激化

- GPT-5: \$11.25/2M tokens
- 中国モデル平均: \$1.50-2.80/2M tokens (87-75%安)
- DeepSeek V3.2はさらに安価 (\$0.20-3.00/2M)

#### 3. トレーニングコストの民主化

- GPT-4o推定: \$100M+
- Kimi K2実績: \$4.6M (約1/22)
- MiniMax-M1実績: \$534K (約1/187)

### 7.2 地政学的インプリケーション

#### 米中AI競争の新局面:

- 中国企業が米国の輸出規制下でも競争力を維持・拡大
- オープンソース戦略によるグローバル開発者コミュニティの獲得
- Airbnbなど米国企業もAlibaba Qwenなど中国モデルを採用

#### OpenAIの戦略的課題:

- CFOが政府の資金支援 (\$1.4T規模) の必要性を示唆
- 中国モデルの性能向上と価格圧力により収益性に懸念
- 2028年まで黒字化見込みなし (\$14B損失/2026年予測)

### 7.3 企業・開発者への示唆

#### 短期的推奨 (2025-2026年)

1. **マルチプロバイダー戦略:** 単一モデル依存を避け、用途別に使い分け
2. **コスト最適化:** 中国モデルで90-95%のフロンティア性能を1/5-1/10のコストで実現
3. **オープンソース活用:** ファインチューニングとカスタマイズの自由度を活用

#### 中長期的トレンド (2026-2027年)

- 1. 性能よりUXへのシフト: ベンチマークではなく実用性が差別化要因に
- 2. マルチモーダルの進化: ERNIE-4.5-VLのような軽量マルチモーダルが増加
- 3. エージェントの実用化: 200-300回のツール呼び出しを実行できるモデルが主流化

## 8. 結論と提言

### 8.1 総合評価

技術的卓越性: Kimi K2 Thinking

- 多くのベンチマークでGPT-5を上回り、オープンソースでありながらフロンティア性能を実現

コストパフォーマンス: MiniMax-M2

- Claude Sonnetの8%のコストで、エージェントタスクにおいて競争力を発揮

革新性: ERNIE-4.5-VL-28B-A3B-Thinking

- 3B活性パラメータでマルチモーダル推論を実現、「Thinking with Images」が画期的

信頼性・エコシステム: GPT-5.1

- 企業統合、サポート体制、安全性で優位。ただし価格とオープン性で劣る

### 8.2 用途別の最終推奨

用途	第1推奨	第2推奨	理由
研究開発	Kimi K2	MiniMax-M2	オープンソース、高性能、低コスト
エンタープライズ本番	GPT-5.1	Kimi K2	サポート体制、実績 vs 性能・コスト
スタートアップ・MVP	MiniMax-M2	Kimi K2	コスト最小化、十分な性能
マルチモーダル応用	ERNIE-4.5-VL	-	唯一の選択肢（他はテキストのみ）
大規模エージェント	Kimi K2	MiniMax-M2	200-300ツール呼び出し、推論深度
コーディング支援	GPT-5.1	Kimi K2	SWE-bench最高 vs 競技プログラミング優位

### 8.3 今後の注目ポイント

- 1. Gemini 3.0のリリース: 2025年11月に限定プレビュー開始、GPT-5.1への対抗
- 2. DeepSeek V4の登場: 中国のもう一つの有力企業、次世代モデル準備中
- 3. GPT-5.1の独立検証: 第三者ベンチマークでの確認待ち
- 4. ERNIE-4.5-VLの実地評価: マルチモーダル性能の実証待ち

### 8.4 日本企業への提言

知財戦略:

- オープンソースモデル利用時のライセンス確認（Modified MIT、Apache 2.0）
- 大規模商用展開時の帰属表示義務（Kimi K2）

## 技術選定:

- PoC段階: 中国オープンソースモデル（コスト効率）
- 本番環境: リスク許容度に応じてGPT-5.1または中国モデル
- マルチモーダル: ERNIE-4.5-VLまたはGemini/GPT-4o Vision

## コスト管理:

- トークン使用量の詳細モニタリング（特にMiniMax-M2は冗長）
- キャッシング戦略の活用（DeepSeek等）
- 複数プロバイダーによるフォールバック体制

---

## 9. 参考資料・出典

### 公式ドキュメント

1. MiniMax GitHub: <https://github.com/MiniMax-AI/MiniMax-M2>
2. Kimi K2 Hugging Face: <https://huggingface.co/MoonshotAI/Kimi-K2-Thinking>
3. ERNIE-4.5-VL Hugging Face: <https://huggingface.co/baidu/ERNIE-4.5-VL-28B-A3B-Thinking>
4. OpenAI GPT-5.1 公式発表: <https://openai.com/index/gpt-5-1-for-developers/>

### ベンチマーク・分析

5. Artificial Analysis Intelligence Index: <https://artificialanalysis.ai/models>
6. VentureBeat - Kimi K2 Thinking分析（2025年11月7日）
7. Simon Willison's Blog - MiniMax M2 & Agent（2025年10月29日）
8. DataCamp - Kimi K2 Thinking Tutorial（2025年11月10日）

### ニュース報道

9. CNBC - Kimi K2トレーニングコスト報道
10. MacRumors - GPT-5.1ローンチ（2025年11月12日）
11. South China Morning Post - MiniMax M2記録達成（2025年10月28日）
12. VentureBeat - ERNIE-4.5-VL発表（2025年11月12日）

---

**レポート作成日:** 2025年11月15日

**調査対象期間:** 2025年10月26日～2025年11月12日

**次回更新予定:** 2025年12月（Gemini 3.0、DeepSeek V4等の情報を追加）

---

本レポートは公開情報に基づき作成されています。ベンチマーク結果は各企業の自己報告値を含むため、第三者による独立検証が望ましい領域があります。実装前に必ず自社環境でのテストを実施して

ください。