2025年第4四半期 生成AIエコシステムにおける技術的転換点：推論効率、エージェント自律性、および産業的適応に関する包括的調査報告書

Gemini

1. エグゼクティブサマリー

2025年10月後半から11月中旬にかけてのAIモデルのリリースラッシュは、単なる性能競争の延長戦ではなく、生成AI技術のパラダイムシフトを決定づけるものとなった。過去数年間、AI開発は「過剰への競争（Race of Excess）」と形容されるような、パラメータ数の増大と計算リソースの無尽蔵な投入に終始していた [1]。しかし、2025年第4四半期に市場に投入された主要モデル――MiniMax-M2、Kimi K2 Thinking、ERNIE-4.5-VL-28B-A3B-Thinking、そしてGPT-5.1――は、明確に異なる方向性を示している。それは、「実質的な知能（Substance）」と「運用効率（Efficiency）」への回帰であり、開発者や企業が直面する「コスト」「速度」「性能」の不可能な三角形（Impossible Triangle）を解決しようとする試みである [2]。

本期間における技術的進歩の核心は、以下の3点に集約される。
第一に、「選択的活性化（Selective Activation）」とMoE（Mixture-of-Experts）アーキテクチャの成熟である。MiniMax-M2やKimi K2 Thinking、ERNIE-4.5-VLに見られるように、数千億から1兆規模のパラメータを擁しながら、推論時にはその数パーセント（30億～300億パラメータ）のみを活性化させる手法が標準化した。これにより、フロンティアクラスの知能を維持しつつ、推論コストとレイテンシを劇的に削減することが可能となった。
第二に、「思考プロセス（System 2 Thinking）」の実装形態の多様化である。OpenAIのGPT-5.1は「適応型推論（Adaptive Reasoning）」により、タスクの難易度に応じて思考時間を動的に調整するユーザー体験重視のアプローチを採用した 3。対照的に、Moonshot AIのKimi K2 Thinkingは、ツール使用と思考プロセスをインターリーブ（交互実行）させることで、数百ステップに及ぶ長期的な自律動作を実現し、特定のベンチマークでGPT-5を凌駕する性能を示した 4。
第三に、「産業特化型マルチモーダリティ」の深化である。BaiduのERNIE-4.5-VLは、汎用的なチャットボットとしての性能よりも、画像内の詳細な推論（ズームイン・アウトによる能動的探索）や産業用チャートの解析能力に特化し、特定のバーティカル領域における圧倒的な効率性を提示している 5。
本報告書では、これらのモデルが提示する技術的革新、ベンチマークにおける詳細な性能比較、そ

してそれが今後のAIエコシステム、特にエージェント開発や産業応用にもたらす深遠な影響について、入手可能な技術文書と検証データに基づき徹底的に分析を行う。

---

2. 技術的背景：パラメータ至上主義からの脱却と「推論の質」への転換

2025年以前のAI開発競争は、主にモデルのサイズ（パラメータ数）とトレーニングデータの規模を競う「スケーリング則」の支配下にあった。しかし、このアプローチは限界費用と推論レイテンシの増大を招き、実社会での応用、特に複雑なタスクを自律的に遂行する「AIエージェント」の実装において大きな障壁となっていた。2025年第4四半期のリリース群は、この課題に対する複数の技術的回答を含んでいる。

2.1 スパース・モデリングとMoEの必然性

「AI開発は過剰のレースと化した」という指摘 [1] は、業界全体の課題認識を反映している。より多くのGPU、より多くのパラメータを追加することは、知能の向上を約束する一方で、運用コストを指数関数的に増加させた。これに対し、MiniMax-M2やKimi K2が採用したアプローチは、巨大な知識ベース（パラメータの総体）を維持しつつ、実行時には「必要な脳細胞」のみを動員するスパース（疎）な活性化戦略である。例えば、Kimi K2は1兆パラメータという圧倒的な規模を持ちながら、推論ごとのアクティブパラメータは320億（32B）に抑えられている [6]。これは、モデルが「博識（巨大な知識）」でありながら「俊敏（高速な推論）」であることを可能にする唯一の解であり、今後の基盤モデル開発のデファクトスタンダードとなることが確実視される。

2.2 「思考」の商品化と適応型計算

推論能力（Reasoning）は、かつてはモデルのパラメータ数に依存する創発的な能力と考えられていたが、現在では明示的に設計・制御される機能へと進化した。GPT-5.1の「Adaptive Reasoning」やERNIEの「Thinking with Images」は、推論リソース（計算時間やトークン数）を、タスクの複雑さに応じて動的に配分する技術である。これは、すべてのクエリに対して最大出力で応答していた従来の無駄を排除し、経済合理性を保ちながら難問を解決するための重要なイノベーションである。

---

3. MiniMax-M2：開発者中心主義とエージェント・エコノミクス

リリース日: 2025年10月26日
開発元: MiniMax AI

3.1 アーキテクチャと設計思想：実用性の追求

MiniMax-M2は、AIモデルとしての「見栄え（Showpiece）」ではなく、開発現場における「実用的な道具」としての地位を確立するために設計された。技術的には、2,300億（230B）パラメータのMoEモデルでありながら、推論時のアクティブパラメータを約100億（10B）に制限している[1]。この「選択的パラメータ活性化（Selective Parameter Activation）」技術は、Qwen3-Nextなどのモデルに見られる最新のトレンドと同様であり、巨大なモデルの知能を、小規模モデルのコストと速度で提供することを可能にしている。

特筆すべきは、開発者が直面する「コスト」「速度」「性能」のトレードオフ、いわゆる「不可能な三角形」への挑戦である[2]。従来、高性能な海外モデル（Claude Sonnet等）は高価で遅く、安価なモデルは複雑なタスクに耐えられなかった。MiniMax M2は、Claude Sonnetのわずか8%の価格で提供され、かつ2倍の推論速度を実現することで、この市場の空隙を埋めている。これは単なる低価格化ではなく、エージェントがタスクを完遂するために数百回の推論を繰り返す現代の開発ワークフローにおいて、経済的な実現可能性を担保する戦略的な価格設定である。

3.2 エージェント特化型機能と統合環境

MiniMax M2の真価は、単発の対話ではなく、統合開発環境（IDE）やエージェントツールとの連携において発揮される。

- ツール統合: Claude Code、Cursor、Cline、Kilo Code、Droidといった主要なAIコーディング支援ツールへの統合が前提とされており、エンドツーエンドの開発ワークフロー（コーディング、実行、デバッグのループ）に最適化されている[2]。
- コンテキスト処理: 128kのコンテキストウィンドウを持ち、複数のファイルにまたがるリファクタリングや、複雑な依存関係の解決において、文脈を失わずに処理を継続する能力が高い。これは、開発者が「散らかったコードのデバッグ」や「大規模なリファクタリング」を行う際に不可欠な能力である[1]。
- 実務での検証: MiniMax社内においても、人事（HR）の履歴書スクリーニングや技術的調査、ユーザーフィードバックの処理といった業務で、M2を搭載したエージェントが人間と協働していることが報告されている[2]。これは、モデルが実験室の中だけでなく、実企業のオペレーションに

耐えうる堅牢性を持っていることを示唆している。

3.3 性能評価と市場での立ち位置

ベンチマークにおいて、MiniMax-M2は「静かなるヘビー級（Quiet Heavyweight）」と評される 7。Artificial Analysisの評価によれば、数学、科学、指示追従、コーディングの総合スコアで、オープンソースモデルとして世界1位を記録したとされる 8。特に、Terminal-BenchやMulti-SWE-Benchスタイルのタスクでの高評価は、同モデルが「コードを書く」だけでなく「環境を操作する」能力に長けていることを裏付けている。
また、画像生成モデル「Nano Banana」との連携など、マルチモーダルな拡張性も視野に入れており、テキスト処理にとどまらない包括的なソリューションプロバイダーとしての側面も持つ 1。

---

4. Kimi K2 Thinking：1兆パラメータの知能と自律的思考の統合

リリース日: 2025年11月6日
開発元: Moonshot AI (中国)

4.1 1兆パラメータMoEと思考プロセス

Kimi K2 Thinkingは、中国のMoonshot AIが投入した野心的なモデルであり、その規模とアーキテクチャにおいて既存の常識を覆すものである。総パラメータ数は1兆（1T）に達するが、推論ごとのアクティブパラメータは320億（32B）に制御されている [6]。この構成は、61層、隠れ層次元7168、そして384のエキスパートからトークンごとに8つを選択するという極めて大規模かつ疎なMoE構造によって実現されている [6]。

Kimi K2の最大の特徴は、モデル名にもある「Thinking（思考）」のプロセスにある。これはOpenAIのo1やGPT-5 Thinkingと同様のアプローチであるが、Kimi K2は「思考（Chain of Thought）」と「ツール使用（Function Calling）」を密接にインターリーブ（交互実行）させる点で独自性を持つ。この設計により、モデルは単に答えを出すだけでなく、自律的にウェブ検索を行い、Pythonコードを実行して検証し、その結果に基づいて次の思考ステップを修正するというループを、人間の介入なしに**200〜300**ステップにわたって継続できる [4]。従来のモデルが30〜50ステップ程度で目標を見失う（ドリフトする）傾向があったのに対し、Kimi K2の長期的な安定性は、自律型エージェントの実用化における

ブレークスルーと言える。

4.2 ベンチマーク分析：GPT-5への挑戦

Kimi K2 Thinkingの性能は、特定の高難易度ベンチマークにおいて、当時の最先端であるGPT-5やClaude Sonnet 4.5を凌駕する数値を記録し、業界に衝撃を与えた。

4.2.1 Humanity's Last Exam (HLE) における躍進

最も注目すべき成果は、数千の専門家レベルの質問からなるクローズドなベンチマーク「Humanity's Last Exam (HLE)」でのパフォーマンスである。

- スコア: ツール使用あり（With Tools）の設定において、Kimi K2 Thinkingは**44.9%**を記録し、GPT-5（41.7%）やClaude Sonnet 4.5 Thinking（32.0%）を上回った [4]。
- 要因分析: ツールなし（No Tools）の設定では23.9%にとどまることから [6]、この勝因は純粋な知識量というよりも、検索やコード実行といった外部ツールを適切に計画・実行し、その結果を統合して解を導く「エージェント能力」の高さにあると分析できる。これは、知識の暗記ではなく、問題解決プロセスの設計能力においてKimi K2が優れていることを示唆している。

4.2.2 数学およびコーディング能力

- **AIME 2025:** 数学ベンチマークにおいて、ツールなしでは94.5%だが、Pythonツールを使用した場合は**99.1%**という驚異的なスコアを記録している [6]。これは、モデルが自身の計算能力の限界を理解し、適切にコードインタプリタを利用して正確な解を導出できることを意味する。
- **SWE-Bench Verified:** 実践的なソフトウェアエンジニアリングタスクにおいては71.3%を記録した。これはGPT-5（74.9%〜76.3%）には及ばないものの、GPT-4o（30.8%）を大きく引き離しており、オープンソースベースのモデルとしてはトップクラスの実力を持つ [6]。
- **BrowseComp:** ウェブブラウジングを伴うタスクでは60.2%を記録し、GPT-5（54.9%）を上回った [13]。これは、最新情報の検索や複雑な情報の統合において、Kimi K2の検索戦略が優れていることを示している。

4.3 インフラストラクチャとアクセシビリティ

1兆パラメータという規模にもかかわらず、Kimi K2は実用性を重視している。ネイティブなINT4量子化と「MuonClip」最適化により、推論レイテンシとメモリ使用量を削減し、一般的な推論環境での展開を可能にしている [14]。また、256kトークンという長大なコンテキストウィンドウは、長編小説の解析や大規模なコードベースの理解において強力な武器となる。

---

5. ERNIE-4.5-VL-28B-A3B-Thinking：産業界の視覚的推論エンジン

リリース日: 2025年11月11日
開発元: Baidu (中国)

5.1 アーキテクチャ：極限までの効率化

Baiduが発表したERNIE-4.5-VL-28B-A3B-Thinkingは、汎用モデルとは異なる独自の進化を遂げている。総パラメータ280億（28B）に対し、推論時のアクティブパラメータはわずか**30億（3B）**という極めて軽量な構成を採用している [5]。この設計思想は、クラウド上の巨大なAPIサーバーではなく、工場のサーバールームやエッジに近い環境での運用を強く意識したものである。80GBのGPUメモリ（シングルカード）で動作可能であることは、多くの企業にとって導入のハードルを劇的に下げる要因となる [5]。

5.2 「Thinking with Images」と強化学習の革新

ERNIE-4.5-VLの最大の特徴は、視覚情報に対する「思考」アプローチにある。

- **Thinking with Images:** 人間が複雑な図面や風景を見る際、全体を漠然と見るのではなく、重要な部分に注目し、細部を確認するために目を近づける。ERNIEはこのプロセスを模倣し、画像を動的にズームイン・ズームアウト（クロッピング）し、細部を反復的に推論する機能を備えている [15]。これにより、高解像度の画像全体を一度に処理する計算コストをかけずに、必要な情報の粒度を確保することができる。
- マルチモーダル強化学習: 学習プロセスには、「GSPO」および「IcePop」と呼ばれる高度な強化学習戦略が導入されている。これらはMoE（Mixture-of-Experts）の学習を安定化させると同時に、「検証可能なタスク（Verifiable Tasks）」——答えが明確に定まる数学や論理パズルなど

——を用いた学習を通じて、モデルの幻覚（ハルシネーション）を抑制し、論理的整合性を高める効果を持つ [5]。また、「Dynamic Difficulty Sampling」により、モデルの学習進度に合わせて難易度の高い例題を重点的に学習させることで、効率的な能力向上を実現している。

5.3 産業ベンチマークにおける優位性

パラメータ数は少ないものの、特定の視覚推論タスクにおいては、GPT-5やGeminiといったフラッグシップモデルを凌駕する性能を示している。

- **ChartQA:** チャートやグラフの解析能力を測るベンチマークにおいて、ERNIEは**87.1**を記録し、Gemini（76.3）やGPT-5（78.2）を大きく引き離している [17]。これは、金融データや生産管理データの可視化解析において、ERNIEが世界最高峰の能力を持つことを意味する。
- **MathVista:** 視覚的な数学問題解決において**82.5**を記録し、Gemini 2.5 Pro（82.3）やGPT-5 High（81.3）を上回った [17]。写真から数式を読み取り、計算を行う能力は、教育支援や研究データの自動入力といった用途で極めて有用である。
- 産業応用: Baiduは、このモデルが物流における「ピークタイムリマインダー」チャートの解析や、回路図におけるオームの法則・キルヒホッフの法則の適用といった、高度に専門的なタスクで検証されていることを強調している [17]。これは、テキストベースのLLMでは到達できなかった「現場の知能化」を実現するものである。

---

6. GPT-5.1：人間中心設計と適応型知能の融合

リリース日: 2025年11月12日
開発元: OpenAI (米国)

6.1 進化の方向性：IQとEQの同時追求

OpenAIがリリースしたGPT-5.1は、GPT-5からのマイナーアップデートという名称でありながら、ユーザー体験とシステム効率の観点から重要な進化を遂げている。その開発の焦点は、単なるベンチマークスコアの向上だけでなく、「人間にとって使いやすいAI」への回帰にある。これまでのモデルが知能（IQ）を追求するあまり「冷たく、機械的」になっていたという批判に対し、GPT-5.1は「より暖かく、会話的」なトーンをデフォルトで採用し、EQ（感情知能）の側面を強化した [3]。

6.2 Adaptive Reasoning（適応型推論）のメカニズム

技術的な最大の革新は、**「Adaptive Reasoning（適応型推論）」**の実装にある。

- 動的なリソース配分: 従来のモデル（GPT-5 Thinking等）は、どのような質問に対しても一定の思考リソースを割く傾向があった。GPT-5.1 Thinkingは、タスクの難易度を即座に判断し、簡単なタスクでは思考時間を短縮して高速に応答し、複雑なタスクではより深く、長く思考するように設計されている。
- トークン生成の変化: OpenAIの内部データによれば、GPT-5と比較して、GPT-5.1 Thinkingは「最も簡単なタスク（10パーセンタイル）」において生成トークン数が**57%**減少し、逆に「最も難しいタスク（90パーセンタイル）」では**71%**増加している [18]。これは、モデルが「手抜き」をしているのではなく、人間の専門家のように「簡単なことは即答し、難しいことは熟考する」という自然な振る舞いを獲得したことを意味する。この結果、全体的なスループットは向上し、ユーザーの待ち時間（Latency）と運用コストの最適化が図られている。
- モードの分化: ユーザーは「Instant（即応・会話重視）」と「Thinking（深考・推論重視）」の2つのモードを選択できるが、「Auto」機能により、クエリの内容に応じて最適なモデルが自動的にルーティングされる仕組みも強化されている [18]。

6.3 安全性とアライメントの強化

GPT-5.1のシステムカード（System Card）には、新たな安全性への取り組みが詳述されている。

- 精神的依存への対策: AIとの対話が自然になるにつれ、ユーザーがAIに過度に依存するリスクが高まる。GPT-5.1では、「精神衛生（Mental Health）」や「感情的依存（Emotional Reliance）」に関する評価指標が新たに導入され、ユーザーの孤立した妄想や不健全な愛着を助長しないよう、出力が調整されている [19]。
- ジェイルブレイク耐性: 「StrongReject」メトリックにおいて、GPT-5.1 Thinkingは0.967、Instantは0.976という高いスコアを記録しており、悪意のあるプロンプトに対する防御力が強化されている [19]。

6.4 ベンチマークと実性能

GPT-5.1は、適応型推論の導入により、一部のベンチマークでスコアの変動が見られるものの、依然として汎用モデルとしての王座を維持している。

- **SWE-Bench Verified:** コーディング能力においては**76.3%**を記録し、GPT-5（72.8%）から

着実な向上を見せた[12]。これにより、複雑なコードベースの修正や機能追加において、より信頼性の高いパートナーとなっている。
- **AIME 2025:** ツールなしでの数学スコアは94.0%であり、GPT-5（94.6%）と比較して誤差範囲の変動である[12]。しかし、これは「効率化」とのトレードオフであり、Thinkingモードを最大限に活用することで、実問題への対応力は向上しているとされる。
- **GPQA Diamond:** 科学的知識を問う難関ベンチマークにおいて88.1%を記録し、前モデル（85.7%）からの進化を示している[12]。

また、「No Reasoning」モード（思考プロセスを省略する設定）においても、ツールの呼び出しやウェブ検索などの低レイテンシが求められるタスクで、従来のGPT-5（Minimal Reasoning）と比較して20%の性能向上が報告されており、API利用者にとってのコストメリットも大きい[20]。

---

7. 比較分析：コスト、性能、そして地政学

2025年第4四半期のAIモデルランドスケープを俯瞰すると、各モデルがターゲットとする市場と強みが明確に分化していることが見て取れる。

7.1 コストパフォーマンスとAPI経済圏

API価格の観点からは、中国発のモデルが破壊的な価格競争力を提示している。

モデル	入力価格 **($/1M tokens)**	出力価格 **($/1M tokens)**	コンテキスト長	特徴・ターゲット
GPT-5.1	$1.25	$10.00	400k (API)	最高品質、汎用性、安全性重視のエンタープライズ
Kimi K2 Thinking	$0.60	$2.50	256k	長文脈のリサーチ、自律エージェント、コーディング

ERNIE-4.5-VL	**$0.14** [21]	**$0.56** [21]	30k-128k	圧倒的な低コスト。産業用画像解析、大量データ処理
MiniMax-M2	(Claude Sonnetの約8%)	-	128k	開発者向けツール統合、コスト重視のエージェント開発

ERNIE-4.5-VLの入力価格はGPT-5.1の約9分の1であり、大量の画像やドキュメントを処理する必要がある産業用途において、圧倒的な経済的優位性を持つ。一方、GPT-5.1はその高価格を「信頼性」「安全性」「多機能性」で正当化しており、プレミアムセグメントを維持している。Kimi K2とMiniMaxは、その中間で「実用的な知能」を安価に提供し、特に開発者コミュニティやスタートアップからの支持を集めようとしている。

7.2 「推論能力」の民主化と分岐

かつてOpenAIの独壇場であった「高度な推論（Reasoning）」は、もはやコモディティ化した。Kimi K2 ThinkingがHLEでGPT-5を上回った事実は、推論技術の覇権が単一の企業に留まらないことを証明している。
しかし、その「推論」の質には違いがある。米国勢（OpenAI, Anthropic）は、あらゆるタスクに対応できる「汎用的な推論」と「安全性」を重視しているのに対し、中国勢（Moonshot, Baidu）は、コーディング、数学、チャート解析といった「測定可能で実用的なタスク」における推論能力を特化させている。Kimi K2の数学・コーディング性能や、ERNIEのチャート解析能力は、この「バーティカル（垂直統合）AI」戦略の現れであり、特定の業務領域においては汎用モデルよりも高いROI（投資対効果）をもたらす可能性がある。

7.3 インフラと展開の柔軟性

展開の柔軟性においてもアプローチが異なる。GPT-5.1は巨大なAPI経由での利用が前提だが、ERNIE-4.5-VLは80GB GPU 1枚での動作を保証しており、オンプレミス環境や閉域網での運用を求める製造業や金融業にとって魅力的な選択肢となる。MiniMaxやKimiも量子化技術を積極的に導入しており、コンシューマーグレードに近いハードウェアでの推論実行を視野に入れている。これは、AIの計算処理をクラウドからエッジ（現場）へと分散させる流れを加速させるものである。

8. 結論と展望：2026年に向けたエージェントの台頭

2025年第4四半期にリリースされたMiniMax-M2、Kimi K2 Thinking、ERNIE-4.5-VL-28B-A3B-Thinking、GPT-5.1は、生成AI技術が「誇大広告（Hype）」の段階を脱し、「実用（Utility）」の段階へ完全に移行したことを示している。

もはや「パラメータ数が最大のモデル」が最良のモデルではない。ユーザーは自身の課題に応じて、最適な「脳」を選択する必要がある。

- **GPT-5.1**は、人間との自然な対話、高度なニュアンスの理解、そして絶対的な安全性が求められる顧客対応やクリエイティブワークにおいて、依然として最良の選択肢である。その「適応型推論」は、ユーザー体験を損なうことなく知能を提供する洗練されたソリューションだ。
- **Kimi K2 Thinking**は、研究開発、複雑なシステム構築、長期間にわたる自律的な調査が必要なシナリオにおいて、その真価を発揮する。特にツールを使いこなす能力は、AIを「チャット相手」から「同僚」へと昇華させる。
- **ERNIE-4.5-VL**は、工場の生産ライン、金融アナリストのデスク、研究室の顕微鏡の横など、視覚データと論理的推論が交差する現場において、比類なき効率性を提供する。
- **MiniMax-M2**は、無数のエージェントが飛び交う未来のソフトウェア開発現場において、そのコストパフォーマンスと速度でインフラとしての役割を果たすだろう。

2026年に向けて、これらのモデルはさらに自律性を高め、人間が介在せずともタスクを完遂する能力を競い合うことになる。我々は今、AIが単に「知っている」だけの存在から、「考え、行動し、結果を出す」存在へと進化する歴史的な転換点を目撃しているのである。

引用文献

1. MiniMax-M2: Better Than GLM 4.6 (Compact & High-Efficiency AI Model) - Analytics Vidhya, 11月 15, 2025にアクセス、https://www.analyticsvidhya.com/blog/2025/10/minimax-m2/
2. MiniMax M2, 11月 15, 2025にアクセス、https://www.minimax.io/news/minimax-m2
3. OpenAI Upgrades GPT-5 With Warmer Tone, Faster Reasoning, and Custom Personality Options, 11月 15, 2025にアクセス、https://www.starkinsider.com/2025/11/openai-upgrades-gpt-5-with-warmer-tone-faster-reasoning-and-custom-personality-options.html
4. Introducing Kimi K2 Thinking, 11月 15, 2025にアクセス、https://moonshotai.github.io/Kimi-K2/thinking.html
5. Baidu launches Ernie-4.5 AI model to rival OpenAI, 11月 15, 2025にアクセス、https://www.techinasia.com/news/baidu-launches-ernie-4-5-ai-model-to-rival-openai
6. My Hands-On Review of Kimi K2 Thinking: The Open-Source AI That's Changing the Game, 11月 15, 2025にアクセス、

https://www.reddit.com/r/LocalLLaMA/comments/1oqi4qp/my_handson_review_of_kimi_k2_thinking_the/
7. MiniMax M2 vs GPT-4o vs Claude 3.5 Benchmark 2025 - - Skywork.ai, 11月 15, 2025にアクセス、https://skywork.ai/blog/llm/minimax-m2-vs-gpt-4o-vs-claude-3-5-benchmark-2025/
8. MiniMax-AI/MiniMax-M2: MiniMax-M2, a model built for Max ... - GitHub, 11月 15, 2025にアクセス、https://github.com/MiniMax-AI/MiniMax-M2
9. Kimi K2 Chinese AI model beats ChatGPT 5 in Humanity's Last Exam, Nvidia CEO says China will win AI race, 11月 15, 2025にアクセス、https://www.indiatoday.in/technology/news/story/kimi-k2-chinese-ai-model-beats-chatgpt-5-in-humanitys-last-exam-nvidia-ceo-says-china-will-win-ai-race-2815836-2025-11-08
10. moonshotai/Kimi-K2-Thinking - Hugging Face, 11月 15, 2025にアクセス、https://huggingface.co/moonshotai/Kimi-K2-Thinking
11. Kimi K2 vs GPT-5 Reasoning: Benchmark Battle & Real Tests ..., 11月 15, 2025にアクセス、https://skywork.ai/blog/agent/kimi-k2-vs-gpt5-reasoning/
12. OpenAI launches GPT-5.1 API with improved coding capabilities and new developer features, 11月 15, 2025にアクセス、https://the-decoder.com/openai-launches-gpt-5-1-api-with-improved-coding-capabilities-and-new-developer-features/
13. Kimi K2 Thinking: Open-Source LLM Guide, Benchmarks, and Tools | DataCamp, 11月 15, 2025にアクセス、https://www.datacamp.com/tutorial/kimi-k2-thinking-guide
14. Kimi K2 Thinking - API, Providers, Stats - OpenRouter, 11月 15, 2025にアクセス、https://openrouter.ai/moonshotai/kimi-k2-thinking
15. Baidu Releases ERNIE-4.5-VL-28B-A3B-Thinking: An Open-Source and Compact Multimodal Reasoning Model Under the ERNIE-4.5 Family - MarkTechPost, 11月 15, 2025にアクセス、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. baidu/ERNIE-4.5-VL-28B-A3B-Thinking released. Curious case.., 11月 15, 2025にアクセス、https://www.reddit.com/r/LocalLLaMA/comments/1ou14ry/baiduernie45vl28ba3bthinking_released_curious_case/
17. Baidu ERNIE multimodal AI beats GPT and Gemini in benchmarks - AI News, 11月 15, 2025にアクセス、https://www.artificialintelligence-news.com/news/baidu-ernie-multimodal-ai-gpt-and-gemini-benchmarks/
18. GPT-5.1: A smarter, more conversational ChatGPT, 11月 15, 2025にアクセス、https://openai.com/index/gpt-5-1/
19. GPT-5.1 Instant and GPT-5.1 Thinking System Card ... - OpenAI, 11月 15, 2025にアクセス、https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_syste

m_card.pdf
20. Introducing GPT-5.1 for developers, 11月 15, 2025にアクセス、https://openai.com/index/gpt-5-1-for-developers/
21. Transparent Pricing for Model APIs & GPU Solutions - Novita AI, 11月 15, 2025にアクセス、https://novita.ai/pricing