ChatGPT Deep Research

[1] [2] **MiniMax-M2の概要**（2025年10月26日リリース） – MiniMax社が開発した2300億パラメータ規模の大規模言語モデル（LLM）で、Mixture-of-Experts (MoE) 型アーキテクチャを採用しています。推論時に実際に活性化するパラメータは100億程度であり、巨大モデルでありながら効率良く動作するのが特徴です [1] [2]。この構造により、従来であれば多数の高性能GPUが必要な処理を**わずか4枚のH100 GPU**で賄うことに成功しており、FP8精度での推論にも対応します [3] [4]。MiniMax-M2はテキスト生成に特化したモデルで、画像や音声といったマルチモーダル機能は備えていません。その代わり、**コーディング、推論、エージェント型タスク**に集中した設計になっており、料金はOpenAIやAnthropicなどの競合他社モデルの約8%という極めて低価格に設定されています [5] [6]（入力100万トークンあたり0.30ドル、出力100万トークンあたり1.20ドル） [7]。さらに**推論速度**も高速で、クラウドAPI経由ではClaude Sonnet 4.5（Anthropic社のモデル）の約2倍のスループットを達成したとされています [8] [9]。こうした**低コスト・高速**という実用上のメリットから、MiniMax-M2は公開直後より注目を集め、リリース時には2025年11月7日まで無料でAPI利用可能なトライアルも提供されました [7]。

性能とベンチマーク – MiniMax-M2は様々なベンチマークで**トップクラスの性能**を示しています。例えば、AI総合能力を測るArtificial Analysis Intelligence Indexではスコア61を記録し、OpenAIのGPT-5（スコア68）に7ポイント差まで迫りました [6]。これは前年における中国産オープンモデルとGPTシリーズの差（18ポイント）から飛躍的な縮小であり、最新モデル間の性能収束が進んでいることを示します [6]。また**コーディング能力**も非常に高く、SWE-Bench（ソフトウェアエンジニアリング課題）のスコアは69.4で、GPT-5の74.9やAnthropic Claude Sonnet 4.5の77.2に肉薄しています [10]。ターミナルでの問題解決を測るTerminal-Benchでは46.3をマークし、Claude 4.5の50.0には及ばないもののGPT-5の43.8を上回りました [10]。特筆すべきは**自律エージェント的な作業**で、ウェブブラウジングエージェントの性能を測るBrowseCompベンチマークでMiniMax-M2は44.0という高スコアを出しています [11]。これはClaudeが19.6でシステムクラッシュさえ起こすタスクにおいて、MiniMaxが安定した遂行能力を発揮したことを意味します [11]。総合的に見て、MiniMax-M2は複数のベンチマークで世界トップ5に入る性能を示し、OpenAIのGPT-5やGoogleのGemini 2.5 Pro、AnthropicのClaude 4.1 Opusといった**最先端閉源モデルに匹敵する**実力を持つことが明らかになりました [6] [12]。

アーキテクチャの特徴と前世代からの向上点 – MiniMax-M2の大きな特徴は、前述のように**MoE（多数の専門家モデル）を用いたハイブリッド構造**で効率化を追求している点です [2]。総パラメータ数は2300億ですが、各トークン予測で用いるのはその一部（~100億）に過ぎません。この「必要な専門家だけ活性化する」アプローチにより、MiniMax-M2は**性能・速度・コストの最適バランス**を実現しています [8] [13]。前世代のMiniMax-M1（2024年リリース）は総パラメータ4560億・活性化パラメータ約459億という大規模モデルで、コンテキスト長80kトークン版も存在するなど野心的でした [14] [15]。しかしM1は高性能な反面、推論コストや速度で課題が残り、GPT-4世代（OpenAI o3等）とのギャップもいくぶん大きいものでした [15] [16]。MiniMax-M2では**活性化パラメータを大幅に削減（~10億）**しつつも新たな学習手法で性能劣化を防ぎ、**前世代比で推論速度50%向上、コスト1/3以下**を達成したと報じられています [17] [18]。さらに**コンテキストウィンドウ**が約200kトークンへと拡張され（M1は128kトークン） [19] [18]、長大な入力（数十万トークン級ドキュメント）にも対応可能になりました。これらの改良により、MiniMax-M2は開発元が当初目標とした「自社内エージェントで本当に使えるモデル」像に一歩近づき、同社は実際にM2を社内の複雑なデータ分析・日常プログラミング・情報検索エージェントに統合して効果を上げ始めています [20] [21]。

MiniMax-M2の得意分野と弱点 – 得意とするタスクは**マルチステップのエージェント実行**と**コーディング**です。複数のツールを順序立てて呼び出しながらタスクを完遂する能力（いわゆる「エージェンティック」な能力）に優れ、シェルコマンド実行・ウェブ検索・Pythonコード実行など外部ツール連携を必要とする長いタスクでも安定してこなします [22]。コード生成に関しては、MiniMax-M2は動作するコードを高い確率で出力し、構造やドキュメントも整ったものを返す傾向があります [23]。特に複雑なデバッグシナリオに強く、長いトラブルシューティングでもコンテキストを保持して問題解決を進めることができます [23]。また**多言語**

対応にも優れ、単なる逐語訳でなく文脈や文化的ニュアンスを踏まえた翻訳が可能だと評価されています [24]。一方、弱点として指摘されているのは**出力の冗長さ（冗長な語り癖）**です [25]。ある評価では、MiniMax-M2は同じタスクを実行するのに競合モデルの3〜4倍のトークン数を消費したとされ [26]、この冗長さはユーザのコスト負担増や応答遅延に繋がり得ます。また高度な**数学的証明や哲学的議論、創造的文章**などではGPT-5やClaudeなど「フロンティアモデル」に一歩譲るとの分析もあります [27]。さらに**マルチモーダル能力を持たない**ため、画像や音声を伴うタスクには対応できません [28]。総じてMiniMax-M2は、**「特定用途に最適化された高効率LLM」**として、汎用性で最新鋭モデルに及ばない部分はあるものの、実用上重要な領域であるコード生成・エージェント実行において驚異的なコストパフォーマンスと高スコアを実現している点が評価できます。

[29] [30] **Kimi K2 Thinkingの概要**（2025年11月6日リリース） – Moonshot AI社による最新のオープンソースLLMで、**「考えるエージェント」**を標榜したモデルです。パラメータ規模はMiniMax-M2を凌ぐ**1兆（1,000B）**に達しますが、こちらもMoEアーキテクチャを採用しており、推論で活性化するのは約320億パラメータに抑えられています [29]。文脈長は256kトークンに及び、極めて長い入力（書籍数冊分の長さ）にも対応可能です [29]。K2 Thinking最大の特徴は、**推論プロセスに「Thinking（思考）」ステップを組み込んでいる**点です。具体的には、モデルが回答を出力する際に**内部チェーン・オブ・ソート（思考プロセス）を逐次トークナイズし、必要に応じて外部ツールを数百回にわたり呼び出しながら推論**を進めます [31]。Moonshot AIは、このモデルが**人間の介入なしに200〜300回ものツールコールを連続実行**できると説明しており [32]、長大な論理推論を伴う問題や情報検索タスクに対して**飛躍的な持久力**を持つことが示唆されています [33]。この「ツール使用と思考のインタリーブ（交互実行）」は、AnthropicのClaudeシリーズが得意とする手法をオープンモデルで実現したものと言えます [34]。実際、ClaudeのThinkingモード同様、K2 Thinkingはツール呼び出しの合間に**<thinking>**などのトークンで始まる内省的な内容（思考過程ログ）を出力する仕組みが実装されています [34]。さらにMoonshot社は**量子化対応学習（QAT）**によってモデルをINT4精度で動作させることに成功しました [30]。これにより**推論速度が約2倍**に向上しつつ、ベンチマーク上の精度低下はほとんど見られないとのことです [30]。報告されているベンチマーク結果はすべてINT4推論で得られたもので、現実的なサービス環境での性能をそのまま示しています [30]。K2 Thinkingは**完全オープンウェイト**でHugging Face上に公開されており、改変・商用利用も可能な緩やかなライセンス（MITベース、極大規模サービスでの表示義務条項付き）で提供されています [35] [36]。ただしモデル重量は約600GBにも達し、自己ホストには相応のGPUメモリ資源が必要です [37] [38]。そのためMoonshot社自身もクラウドAPIを提供しており、API利用料金は入力100万トークンあたり0.60ドル・出力100万トークンあたり2.50ドルとアナウンスされています [39]。これはOpenAIやAnthropicと比べ極めて競争力のある価格設定です（前述のMiniMax-M2よりは高いものの、それでもGPT-5の約10分の1程度） [39] [7]。

性能とベンチマーク – K2 Thinkingの登場はAI研究コミュニティに衝撃を与えました。というのも、**いくつかの難関ベンチマークでK2がGPT-5（OpenAIの当時最新モデル）を凌駕した**と発表されたからです。代表的なのは**Humanity's Last Exam (HLE)**と呼ばれる総合推論テストで、K2は44.9%という当時の最高成績を収めています [40]。このスコアは、GPT-5（Thinkingモード搭載版）やAnthropic Claude Sonnet 4.5といった閉源モデルを上回るもので、**HLEにおける新たな世界記録**となりました [40]。また**BrowseComp**（長文読解・ウェブ検索エージェント評価）でも60.2%をマークし、こちらもGPT-5 (54.9%)やClaude 4.5 (24.1%)を引き離しています [41] [42]。コーディング関連では、前述のSWE-Bench Verifiedで71.3%というトップ級スコアを達成し [40]、LiveCodeBench v6（動的実行テスト付きのコード生成評価）でも83.1%と高い正答率を示しました [43]。さらに検索ベースの情報取得ベンチマークであるSEAL-0でも56.3%を記録しており [43]、総じて**推論・エージェント・コーディングの各分野で業界最高水準**にあることがわかります。Moonshot社の発表によれば、K2 Thinkingは**AI総合指数（Artificial Analysis Intelligence Index）でスコア67**を獲得しており、これはオープンウェイトモデル中で1位、全モデル中でもGPT-5 (68)に次ぐ第2位だったとのことです [44]。一部のテストではGPT-5を僅差で下回る場合もありますが（例えば数学コンテストAIME 2025ではGPT-5が94.6%、K2がほぼ同等の94.0% [45]）、**総合的にはK2は多くの項目でGPT-5と同等以上**に達しています。Carl Franzen氏による報道では、GPT-5とK2の比較として「**GPQA Diamond**（高度質問応答タスク）でK2が85.7%対GPT-5の84.5%と上回り、数学系のAIME 2025やHMMT 2025でもほぼ肩を並べる」と評価されています [41] [46]。またK2はAnthropic Claudeシリーズにも匹敵・一部凌駕し、特に長い思考を要するタスクでは

Claudeを大きくリードしています（例：BrowseCompでClaude 4.5の24.1%に対しK2は60.2%） [41] 。要するにK2 Thinkingは、**「オープンソースモデルがGPT-5等の閉源トップモデルに初めて肩を並べた」**と表現し得る画期的な存在となったのです [47] [48] 。実際、Moonshot社はこの成果を受けて「閉源のフロンティアとオープンモデルの性能差が事実上ゼロに近づいた」とコメントしており [49] 、コミュニティでも「オープンウェイトの勝利」として話題になりました。

他モデルとの比較とK2の位置付け – K2 Thinkingはリリース直前までオープン界隈で最高性能と称されていたMiniMax-M2を追い抜き、新たな**オープンソースLLMの王者**となりました [50] 。例えば、MiniMax-M2のBrowseCompスコア44.0に対しK2は60.2を記録し、SWE-BenchでもK2の71.3%がM2の69.4%を上回っています [51] 。金融分野の推論（FinSearchComp）ではMiniMax-M2が一部優位なケースもあるものの、K2は総合的な汎用推論能力で勝っていると評されています [51] [52] 。技術的にも、両者は共にMoEによる効率化を図っていますが、K2は**より多くのエキスパートを活性化しつつINT4量子化**という先進手法を投入した点で進んでいます [53] 。これが長大な思考過程（256kコンテキストに及ぶようなプロンプト）でも速度・精度を両立するポイントとなりました [53] 。閉源モデルとの比較では、**OpenAI GPT-5**および**Anthropic Claude Sonnet 4.5**（いずれもThinkingモード）という両巨頭に対し、K2は一部ベンチマークでリードを取りました [48] 。GPT-5やClaudeが依然として総合的な知識量や安全性などで優位性を保つ部分もありますが、**K2は推論過程の透明性**（推論コンテンツを逐次出力する機能）や**自律エージェント実行**において極めて洗練されており、差別化に成功しています [54] [55] 。Moonshot社のデモでは、K2が日時取得APIやウェブ検索を自発的に呼び出し、得られた情報を解析してニュースレポートを作成する一連の処理を人手なしで完了する様子が示されました [55] 。このようにK2は**高度に自律したAIエージェント**として動作できることから、実世界での複雑業務（長大な問い合わせへの回答、自動調査・分析など）に適したモデルと位置付けられます。一方で**課題や制約**もいくつか指摘されています。まず出力が詳細になる傾向、すなわち冗長性については、先述のMiniMax-M2同様にK2もかなりの**トークン数を費やして丁寧に推論**する傾向があります [37] 。実験では、同一タスクでGPT-5（Thinking）に比べ約2倍のトークンを消費したと報告され、応答に時間がかかったりコスト増につながったりする可能性があります [37] 。またモデルデータが600GBと巨大なため**セルフホストは難易度が高く**、事実上クラウド上の提供元APIや軽量化サービスに頼るユーザも多いようです [38] [56] （リリース直後にはOpenRouter等のホスティング環境でレイテンシやエラーばらつきが見られたとの報告があります [56] ）。しかしこれらを差し引いても、K2 Thinkingが**「OpenAIやAnthropicの牙城に食い込んだ」**意義は大きく、2025年末時点で最も注目すべきモデルの一つとなっています [47] [48] 。

[57] [58] **ERNIE-4.5-VL-28B-A3B-Thinkingの概要**（2025年11月11日リリース） – Baidu（百度）社が公開した最新世代のマルチモーダルAIモデルです。ERNIEシリーズはこれまでも中国語圏を中心に強力な言語モデルとして知られてきましたが、バージョン4.5-VL-28B-A3B-Thinking（以下、ERNIE-4.5-VL）は**画像と言語の統合処理**と**高度な推論能力**を兼ね備えた特殊なモデルとなっています [59] 。モデル名に含まれる「28B」は総パラメータ数約280億を意味し、「A3B」はおそらく「Active 3B」（推論時活性化パラメータ30億）の略で、MiniMaxやK2と同様**MoE的手法**で効率化が図られていることを示唆します [60] 。実際、ERNIE-4.5-VLは「ライトウェイト」（軽量）モデルと称されており、**わずか30億のパラメータを動かすだけで業界トップクラスの性能を実現した**と公式に謳われています [59] [61] 。Baiduは本モデルをApache 2.0ライセンスで**完全オープンソース**公開しており、商用利用も可能です [62] 。重みデータはHugging Face上で提供されており、Transformersや社内ツール（ERNIEKit）での推論・微調整もサポートされています [62] [63] 。ただし推論には少なくとも1枚あたり80GBメモリ級のGPU（例: NVIDIA A100 80GB）が必要で、単体ではカジュアルな利用に向かないヘビー級モデルでもあります [62] 。

マルチモーダル強化学習と「考える画像処理」 – ERNIE-4.5-VL最大の特徴は、**視覚（画像・動画）とテキストの融合能力**です。Baiduの解説によれば、本モデルは大規模な視覚と言語のペアデータで中間学習（mid-training）され、画像とテキスト間のセマンティック（意味）アライメントを飛躍的に高めたといいます [57] 。さらに**マルチモーダル強化学習**を導入しており、モデルに「検証可能なタスク」を与えて報酬学習を行う際に、**GSPO**（Generalized Self-Play Optimization）や**IcePop**といった新しい手法を組み合わせてMoE訓練の安定化と学習効率向上を達成しています [58] 。これに動的難易度サンプリングも組み合わせ、難易度の高い視覚推論タスクでも効率的に学習できるよう工夫されています [58] 。その成果として、**「Thinking with**

Images」というユニークな機能が挙げられます [64]。ERNIE-4.5-VLはまるで人間のように**画像を自由にズームイン・ズームアウトし、細部まで観察して隠れた情報を発見**することができます [64] [65]。また外部ツールとの連携にも優れており、例えば画像中の不明物体を認識する際には**自動で画像検索ツールを呼び出し**、関連知識を取得してから回答を生成するといった挙動も可能です [66] [67]。このように**視覚情報を積極的に活用・操作して推論**できる点が、従来の画像キャプション生成モデルとは一線を画すポイントです [60] [64]。Baiduはこれを「人間のように画像で考える（Thinking with Images）」能力と称しています [64]。さらに動画にも対応しており、シーンの時間的変化やイベントを捉えて分析する**動画理解**能力も備えているとされています [67]。実際、企業ユースを想定した機能として、動画から全字幕を抽出してタイムスタンプにマッピングする、映像アーカイブから特定シーン（例：「橋の上で撮影された場面」）だけを検索で見つけ出す、といったデモンストレーションが紹介されています [68] [69]。

性能とベンチマーク – ERNIE-4.5-VLは「軽量モデルでありながらフラッグシップ並みの性能」と謳われており [61]、実際に公開されたベンチマークでも**既存の大規模マルチモーダルモデルに匹敵、あるいは凌駕する**結果が示されています。Baiduの社内評価によると、例えば以下のような**視覚と言語の複合タスク**で高成績を記録しています [70]：

- **MathVista**（画像化された数学問題の解答精度）: ERNIE 82.5点 vs Google Gemini 2.5 Pro 82.3点 vs GPT（GPT-5と思われる） 81.3点 [70]。わずかながらERNIEがトップで、数学的図表問題の解釈・解答で世界最高水準にあります。
- **ChartQA**（グラフ・チャートを読み取って質疑応答）: ERNIE 87.1点 vs Gemini 76.3点 vs GPT 78.2点 [71]。統計グラフの理解に関する難関タスクで、ERNIEが他モデルを**大きくリード**しています。
- **VLMs Are Blind**（マルチモーダル文脈理解テストの一種）: ERNIE 77.3点 vs Gemini 76.5点 vs GPT 69.6点 [71]。詳細は不明ですが、名前から察するに「視覚言語モデルでも見落とすような情報を問う」ベンチマークでしょう。ここでもERNIEが最高スコアを達成しました。

これらはBaidu側の発表ですが、第三者メディアも「ERNIEがGPTやGeminiをベンチマークで打ち負かした」と報じています [72]。特にテキスト以外のデータを扱う能力（例えばエンジニアリングの設計図、工場のカメラ映像、医用画像、物流ダッシュボードなど）に着目すると、ERNIE-4.5-VLは**従来テキスト中心のLLMが不得意としていた領域で真価を発揮**します [73] [74]。記事によれば、ERNIEはある実験で「ピーク時刻リマインダー」と題された混雑状況チャートを読み取り、最適な訪問時間帯を見出すという課題をこなしたり、橋回路の回路図を解析してオームの法則・キルヒホッフの法則を当てはめて解を導いたりしたそうです [75]。これは、従来であれば人間の専門知識が必要な**複雑な図面・チャート分析**をAIが自動で行えていることを意味します。さらに、企業内の大量映像データの活用例として、研修や会議の録画から全発話テキストを書き起こして索引化したり、何時間ものセキュリティ映像から特定の状況だけ抜き出したりするデモが紹介されました [68] [69]。要するに、ERNIE-4.5-VLは**「見て、読んで、動く」**マルチモーダルAIとして、これまでテキスト偏重だった生成AIの用途を大きく広げるポテンシャルを持っています [76] [77]。

実用面の仕様と位置付け – ERNIE-4.5-VLは完全オープンソースで公開されたこともあり、**企業が自社データに特化したマルチモーダルAIを構築するための土台**として注目されています。Baiduは開発者向けにTransformers（Hugging Face）やvLLM、FastDeployなど複数のデプロイ手法を提示し [62]、高性能インフラを持つ組織であれば柔軟にこのモデルを組み込めるよう配慮しています。前述の通り単機での動作には80GB GPUが必要で、これは例えばNVIDIA A100 80GBカード1枚に相当します [62]。もし24GB GPUで運用するなら4枚程度が必要になるでしょう。そのため「遊び半分で試すツール」ではなく、**本腰を入れて導入するエンタープライズ向けAI**と位置付けられます [62]。もっとも、オープンライセンスかつ自前データでの微調整も可能なため、要件を満たす企業にとっては極めて有用です。Baidu側も、カスタム用途の微調整（ファインチューニング）は多くのケースで必要になるだろうとして、自社ツールERNIEKitでその作業を支援しています [62]。対応言語について明確な記述はありませんが、Baidu製である以上**中国語**について最適化されているのは確実で、加えて英語や主要言語についても大規模データで学習していると考えられます（実際、上記ベンチマークは英語主体です）。総合すると、ERNIE-4.5-VL-28B-A3B-Thinkingは**「スリムな計算量で重量級モデル並みの賢さを発揮する、実用志向のマルチモーダルLLM」**と評することができます [60] [78]。OpenAIの

GPT-5シリーズやGoogleのGeminiシリーズがリードしてきた視覚と言語の融合分野で、Baiduがオープンかつ高性能な解を提示した点は注目すべきトレンドです。

[79] [80] **GPT-5.1の概要**（2025年11月12日リリース） – OpenAIがGPT-5のアップグレード版としてリリースした最新モデルです。ChatGPT製品への実装を念頭に置いた**対話特化型LLM**であり、大きな特徴として**「GPT-5.1 Instant」**と**「GPT-5.1 Thinking」**という2種類の動作モード（あるいはサブモデル）が同時に提供されました [79]。Instantは従来のChatGPT（GPT-5ベース）の後継で、**より「暖かみがあり、知的で、指示に従いやすい」**会話スタイルへとチューニングされています [81]。一方、Thinkingは高度な推論タスク向けのモデルで、**シンプルな質問にはより素早く、複雑な問題にはより粘り強く**答えるよう調整されています [81]。GPT-4世代から導入された「システムメッセージによるスタイル制御」の発展形として、GPT-5.1では**ユーザがChatGPTの口調やキャラクターを直感的にカスタマイズできる**新機能も発表されました [82]（具体的には8種類のパーソナリティプリセットやスライダー調整機能が追加されたと報じられています [83] [84]）。これにより、利用者の好みやコンテキストに応じて応答のトーンを柔軟に変えられるようになっています。

性能と改善点（GPT-5との比較） – OpenAIはGPT-5.1で**推論能力と応答の質を両立して向上**させることに注力しました。技術的な目玉は、**「適応的推論（adaptive reasoning)」**と呼ばれるメカニズムです [80] [85]。これは、質問の難易度に応じてモデルが内部で思考（チェーン・オブ・ソート）にかけるステップ数や計算資源を調整する仕組みです [80] [85]。簡単なタスクでは無駄な反復を省き、一問一答をほぼ即座に返せるようにし、難しいタスクではこれまで以上に深く検討するようになりました [85]。具体的な効果として、OpenAIのテストでは「日常的な簡易質問に対し、GPT-5.1はGPT-5の5分の1程度の思考トークンで回答できた」と報告されています [86] [87]。例えば「グローバルにインストールされたnpmパッケージの一覧を表示するコマンドは？」という質問では、GPT-5では約10秒（~250トークンの思考）かかったところを、GPT-5.1は2秒（~50トークン）で正答を返したといいます [86] [87]。このように**不要な遅延を削減しつつ、必要な場合は徹底的に考え抜く**という柔軟性がGPT-5.1 Thinkingの特徴です。さらに**“no reasoning”モード**と呼ばれる設定も新設され、ユーザが望むなら明示的に「思考プロセスをスキップして即答してほしい」とモデルに指示でき、応答速度を稼ぐことも可能です [88]。

また**推論効率**の改善として、OpenAIは**24時間有効なプロンプトキャッシュ**機能を強化しました [89]。同一ユーザから連続する問い合わせがある場合、前回の推論内容をキャッシュしておき、再利用可能な部分は90%コストカットされた料金で処理されます [90] [91]。これにより、チャットの文脈維持コストを下げつつ応答を高速化できます。加えて、OpenAIは優先実行権（Priority Processing）を持つ顧客向けに、GPT-5.1では**GPT-5比2〜3倍のスループット向上**を実現したとも述べています [92]。この主張はパートナー企業による評価で裏付けられており、ある金融企業は「GPT-5.1は当社の全動的評価スイートでGPT-4.1とGPT-5の両方を上回り、速度はGPT-5の2〜3倍速かった」とフィードバックしています [92]。また別の保険業AI企業も「エージェントの処理がGPT-5.1で50%高速化し、精度もGPT-5および他の主要モデルより高かった」と報告しました [93]。こうした第三者からの評価も交え、OpenAIはGPT-5.1が**「GPT-5シリーズの次なる進化」であり、より知的かつ迅速なエージェント開発を可能にする**モデルだと位置づけています [94] [95]。

ベンチマーク上の評価も公表されており、GPT-5.1 (high設定) はGPT-5 (high設定) と比較して多くの指標でわずかながら改善を見せています [96] [45]。例えば、SWE-Bench Verified（コーディング問題）では76.3%対72.8%と約3.5ポイント向上し [96]、GPQA Diamond（知識推論）も88.1%対85.7%と2.4ポイント向上しました [97]。MMLU（学術知識テストに類似）では85.4%対84.2%と1ポイント強の改善です [98]。一方でAIME 2025（数学コンテスト）では94.0%対94.6%とほぼ同等、Tau-ベンチ（ツール使用推論）ではシナリオによって微増微減が見られました [45] [99]（例：航空会社シナリオで5ポイント上昇、テレコムシナリオで1ポイント低下 [100] [99]）。長文読解系のBrowseComp（128kトークン長文）では90.0%で両モデル同率でした [101]。総じて劇的な性能向上ではないものの、GPT-5.1はGPT-5に比べて**堅実なブラッシュアップ**が図られていることがわかります。特にコード生成においては、OpenAIがCursorやWarpといった開発者ツール企業と連携してモデルの「コーディング人格」を改善したと述べており [102]、ユーザの指示に対する応答の洗練、コードの一貫性やエラーハンドリングの向上が期待できます。さらに**エージェント機能**として、GPT-5.1 APIには

新たに`apply_patch`ツール（提案コード差分を既存コードに適用する機能）と`shell`ツール（シェルコマンド実行機能）が追加されました[94]。これにより、ChatGPTやAPIを使った自動コード修正・スクリプト実行といったワークフローが一層シームレスになります。

提供状況と実用面 – GPT-5.1はリリース当日からChatGPTのPlusユーザやEnterpriseユーザに順次適用され、APIでも`gpt-5.1`エンジンが利用可能となりました[103][104]。料金設定は大きな変更は公表されていないため、GPT-5と同等か多少割安程度と考えられます（OpenAIは2024年以降大規模モデルの料金を段階的に引き下げてきた経緯があります）。プロンプトキャッシュ利用時には入力トークン料金が90%引きとなる点が新しいトピックでした[90]。モデルの**対応言語**については特筆すべき新規情報はないものの、GPT-4以来OpenAIモデルは主要言語で高い性能を持つことが知られており、GPT-5.xでも英語はもちろん日本語を含む多言語でトップクラスの結果を残していると推測されます（実際、GPT-4は多言語のMMLUテストで上位でした）。GPT-5.1もその延長線上にあり、対話の文脈ではユーザの言語で自然に応答できるマルチリンガルモデルです。なお、GPT-5シリーズの**マルチモーダル対応**に関して明示的な発表はありませんでしたが、ChatGPTにはGPT-4.5世代から画像入力機能が搭載されており、GPT-5.1でもその機能は継承されていると考えられます。ただしリリースノートの焦点は主に対話品質と推論効率であり、画像や音声について大きな変更は言及されていません。総合すると、GPT-5.1は**「より賢く親しみやすいChatGPT」**として位置づけられ、既存ユーザからのフィードバック（GPT-5はやや冗長・生硬だったとの声も一部ありました）を踏まえた改良が随所に見られるアップデートとなっています[84][105]。

4モデルの比較：得意分野と特性 – 以上の4つのモデル（MiniMax-M2, Kimi K2 Thinking, ERNIE-4.5-VL, GPT-5.1）は、それぞれが最先端の性能を持ちながらも**目指す方向や強み**に違いがあります。以下、いくつかの観点で比較します。

- **コーディング能力**: いずれのモデルもコード生成で高性能を示していますが、特に**GPT-5.1**と**Kimi K2**はトップクラスです。K2 ThinkingはSWE-Bench Verifiedで71.3%を出し、MiniMax-M2の69.4%を上回りました[51]。GPT-5.1もGPT-5より精度を上げて76.3%に達しており[96]、現在のベンチマーク上では僅差ながらOpenAIモデルがリードしているように見えます（AnthropicのClaude Sonnet 4.5は77.2%[10]で依然トップですが非オープン）。**MiniMax-M2**はコード出力の実用性に定評があります。構文の正確さやコメントの適切さ、複雑なデバッグの対応力など、実際の開発で役立つ品質を備えています[106]。M2はやや保守的ではあるものの安定したコードを返し、長時間のコーディングセッションでも整合性を保つ点が評価ポイントです[106]。しかも**圧倒的な低コスト**でこれを実現しており、コーディング用途に大量にトークンを消費する場合の経済性では群を抜きます[5][6]。一方、**ERNIE-4.5-VL**は視覚要素を含むタスク（例えば画像中のテーブルデータからコード生成など特殊ケース）を除けば、コード専用モデルほどの評価は出ていません。総合的には、日常的なコーディングではOpenAI（GPT-5.1）の信頼性・使いやすさ、あるいはAnthropic Claude系の一貫性が強みとなり、複雑なマルチファイル編集や自律デバッグではMiniMax-M2やK2のエージェント能力が光る、といった住み分けが考えられます。

- **論理推論・エージェントタスク**: **Kimi K2 Thinking**はこの分野で際立っています。HLE（人類最後の試験）やBrowseCompといった**総合推論系ベンチマークで世界トップ**に立ったように[41]、未知の難問を論理立てて解決する力、そして必要に応じてWeb検索や電卓など外部ツールを組み合わせる力で、K2は他を一歩リードしています。特にBrowseCompではK2が60.2%を取り、GPT-5 (54.9%)やClaude 4.5 (24.1%)より大幅に高いスコアでした[41][42]。これは**オープンモデルが初めて閉源の最先端モデルを凌駕した指標**とも言われています[47]。GPT-5.1も**Thinkingモード**を備え、チェーン・オブ・ソート＋ツール使用能力を強化しています[80][85]。GPTシリーズは元々知識量や基本的論理力で定評があり、K2に負けず非常に強力ですが、**違いが出るのは持久力と透明性**です。K2は何百ステップにもわたる長い推論を人手なしで続行でき、かつ各ステップの思考内容を露わにするため（専門家が途中経過を検証しやすい）[54]、信頼性と汎用性の面でユニークな強みを発揮します。Claude 4.5も思考過程を持つモデルですが、K2のようなツール連携の高度さはなく、長すぎる推論では破綻するケースが報告されています[107]。**MiniMax-M2**もエージェント的タスクが得意で、Tauベンチマーク

（Tool Use試験）でもスコア77.2を記録するなど高評価でした [108]。ただK2の登場で、同じオープンエージェントでもより高性能な選択肢が出てきた形です。M2は現状、K2ほど長大な推論シーケンスには強くないものの、**高速かつ安価にエージェント機能を実装**したい場合に有力な選択肢です。**ERNIE-4.5-VL**に関しては、エージェントというより**マルチモーダルな論理推論**が専門です。テキスト＋画像＋必要に応じWeb検索といった複合タスクでは、ERNIEが他モデルでは真似できない活躍をするでしょう。逆にテキストのみの純粋な論証問題では、K2やGPT-5系列の方が実績上は高スコアです。ただERNIEは設計上、**視覚情報を手がかりに論理推論を行う**能力が突出しており、これらはGPT-5やK2には無い領域です（例：回路図を読み解いて物理法則を適用するなど [109] [110]）。総じて、純テキスト論理ではK2・GPT-5.1が双璧、**マルチモーダル論理**ではERNIE-4.5-VLが一歩抜きん出ていると言えるでしょう。

- **創造性・文章生成**: **GPT-5.1 Instant**は“より暖かく遊び心のある”応答をデフォルトで返すよう訓練されており [111] [112]、ChatGPTとしてのユーザ体験重視の改良が行われました。ユーザの曖昧な指示からでも適切に意図を汲み、堅苦しすぎずフレンドリーなトーンで答える能力は、依然OpenAIモデルの強みです。加えて指示遵守性も向上し、細かなニュアンスの要望にも応えやすくなっています [81]。**Kimi K2 Thinking**も、オープンモデルとしては珍しく**文章表現の質**が称賛されています。第三者によるレビューでは、K2の文体は「定型的でなく自然で、ときに文学的・情感的ですらある」と評されており [113]、多くのLLMが苦手とする**感情豊かなクリエイティブライティング**で新境地を開いている可能性があります。この背景には、K2の前身であるKimi K2 Instructモデルが持っていた洗練された文体をRL訓練後も保持するよう調整されたことがあるようです [114]。Claude 4.5は以前から「文章の上手さ」では定評がありましたが、K2もオープンながら非常に**読みやすく魅力的なテキスト**を生成できるという評が出ています [114]。**MiniMax-M2**は創造的文章より実務的応答を重視したモデルで、簡潔で自己完結的な回答をする傾向があります [23]。そのため、小説執筆や詩作などではGPTやClaudeほど表現力は高くないものの、技術文書や指示の明確な文章では的確かつ過不足ない出力を返します [115]。**ERNIE-4.5-VL**の文章生成能力は視覚コンテキストに大きく依存します。例えば画像の詳細描写や、画像内容を踏まえた創造的ストーリー生成などでは独自の強みを発揮するでしょう。実際、ERNIEは「画像を単に文章に装飾として添えるのでなく、**調査対象として深掘りする**」アプローチを採っています [116]。したがって、図版付き記事の自動生成や画像からインスピレーションを得たコピーライティングといった応用が考えられます。総じて、**汎用クリエイティビティ**ではGPT-5.1が依然リードし、K2がそれに迫る勢い、MiniMax-M2は実用重視、ERNIE-4.5は視覚文脈下で特化した創造性を見せる、と整理できます。

- **マルチモーダル処理**: 4モデル中、この点で際立っているのは明らかに**ERNIE-4.5-VL-28B-A3B-Thinking**です。これは本質的に**視覚と言語の統合AI**であり、入力として画像や映像を直接与え、そこに対する高度な質問応答・分析を行えます [117] [77]。上記のようにERNIEはチャート解析や画像中の文字読み取り、物体検出からの構造化出力までこなし [118] [119]、さらに動画内イベント検出なども可能です [120]。OpenAIのGPT-5.1（およびGPT-5）もChatGPT経由で**画像入力**を受け付ける機能自体はありますが、その能力は主に画像の内容説明や簡易な質問への回答に留まります。GPT-4（Vision）は静止画の説明や問題解決で一定の成果を見せましたが、**ツールを伴う画像推論**や**複数画像の比較**などは現状不得手です。GPT-5.1がその点で飛躍したという情報はありませんので、マルチモーダル性能ではERNIEが先端を行くと考えられます。**Kimi K2**と**MiniMax-M2**はテキスト専用モデルであり、画像や音声を直接扱う機能は持ちません [28]。ただしK2はテキスト上での画像情報検索や想像力を働かせることは可能でしょう（例えば「<画像検索ツール>」を使って架空の結果を得て推論するなどのシミュレーション）。しかし実画像を理解することはできないため、現実世界のマルチモーダルタスクでは**ERNIE-4.5-VL**が唯一の選択肢となります。GoogleのGeminiやOpenAIの将来モデルもマルチモーダル対応を強化していますが、**現時点（2025年11月）でオープンに使える最強の視覚言語モデルは ERNIE-4.5-VL**であると言って差し支えないでしょう [121] [122]。

- **コンテキスト長（メモリ）**: コンテキストウィンドウの長さは、エージェントや長文読解タスクにおいてモデルの実用性を左右します。**Kimi K2 Thinking**は256kという飛び抜けた長コンテキストをサポー

トし[29]、これは現在商用利用可能なモデルの中でも最大級です（Anthropic Claude 3.5が100k超を扱えますが、K2はそれを上回ります）。**MiniMax-M2**も200k超（約20万トークン）を扱えると報じられており[19][18]、こちらも非常に大きいです。実際、K2とM2はいずれも長大なドキュメントを分割せず一度に読み込んで処理するデモが行われています。例えばあるテストでは、Claude 3.5・GPT-4 OpenAI版・MiniMax-M2にそれぞれ12万トークンのリサーチ資料を読ませ要約させたところ、Claudeは安定して引用を交えつつ回答でき、M2とGPT-4 OpenAI版（32k上限）はそれなりに健闘したものの若干強引な要約になる部分もあった、という指摘があります[107][123]。Claudeは長い文脈での挙動が落ち着いている（ライブラリアン的だ）との評価があり[124]、極端な長文処理では現在もClaudeシリーズが一日の長あるようです。しかしK2やM2はそれに迫る容量を持ち、今後チューニングでさらに安定性を増す可能性があります。**GPT-5.1**は社内テストで128kトークンの長文読み取りを行っており、その結果BrowseComp Long ContextではGPT-5と同等の90.0%を出せています[101]。したがってGPT-5系列もClaude並み（ないしそれ以上）の長コンテキスト対応を果たしたと見られます。OpenAIが公式にGPT-5の最大コンテキスト長を公表していませんが、少なくとも128kは扱えることになります。**ERNIE-4.5-VL**については、テキストのみのコンテキスト長はそれほど強調されていません。しかし、画像（高解像度の場合ピクセル情報自体は膨大）や長時間の動画をハンドリングする関係上、別の形で長い「文脈」を処理できるはずです。Baiduの資料では具体値は出ていないものの、例えば数十分の動画から字幕を抽出・検索したりできるので、間接的には数万トークン相当のテキストを扱えるでしょう[68][69]。以上をまとめると、**256kのK2が最大、MiniMax-M2とGPT-5.1/Claude-3.5が約100k〜200k級、ERNIEは形式が違うが大規模視覚文脈OK**という序列になります。長大なドキュメントの分析や複数ドキュメントの一括要約といったニーズには、これら長コンテキスト対応モデルが不可欠です。ただし実際には、人間にとって扱いきれない情報量をモデルが一度に読めても、回答が冗長になったり重要点を見逃したりするリスクもあります。したがって、コンテキスト長競争は単に伸ばせば良いというものではなく、**長文中の要点抽出と整合性維持**という質的課題とセットです。この点、Claudeは安定感で知られ[124]、OpenAIや中国勢がそこをどう改善していくかが今後の焦点と言えます。

- **推論速度（レイテンシ）**: 推論速度はモデルアーキテクチャと最適化技術に大きく依存します。**MiniMax-M2**はMoE活性化パラメータを抑えた効果で**高速なストリーミング**が可能です。同等性能の他モデルに比べ、初回トークン出力までの待ち時間が短いとの報告があり、ある比較ではM2が平均0.9秒で応答を開始したのに対し、GPT-4系列は1.8秒、Claude 3.5はキューイングの影響もあり平均1.8秒だったと言います[9]（テスト条件によるので一概には言えませんが、M2は体感上「待たされにくい」モデルとして好評です）。**Kimi K2**はパラメータ数こそ多いもののINT4量子化で計算を効率化したため、標準精度モデルの約2倍速で生成できます[30]。Moonshot社は**K2を4bit精度で事後学習（QAT）**することで推論を高速化し、長い思考セッションでもレスポンスを改善したと述べています[30]。ただしK2は事例によっては非常に多くのステップを踏むため、**簡単な質問ではむしろ考えすぎて遅く感じる**場合もあるかもしれません（この点GPT-5.1は思考省略モードを用意しています[88]）。**GPT-5.1**はOpenAIが速度面で力を入れたモデルです。特にInstant版はChatGPTでの応答体感速度向上が目的で、複雑でない質問にはほぼ即答に近いレスポンスを返します[86]。Thinking版も従来比では高速化されており[92]、OpenAIによればGPT-5比で2〜3倍高速（同程度の出力内容の場合）とのことです[92]。**ERNIE-4.5-VL**はActive 3Bということで1トークンあたりの計算は軽いですが、画像処理など前処理・後処理部分が含まれるため一概に速いとは言えません。しかし、競合するGPT-4 Vision等と比べれば軽量であり、複数の推論エンジン（vLLMやFastDeploy）で**高スループット動作**を狙える柔軟性があります[62]。注意すべきは、全モデルとも**出力トークン数が増えればそれだけ時間はかかる**点です。MiniMax-M2やK2は冗長な回答をする傾向があるため[25][37]、結果的にユーザが答えを得るまでの待ち時間が伸びる可能性があります。一方GPT-5.1やClaudeは比較的簡潔にまとめる傾向があります（もっとも指示次第ですが）。このように、**速度 vs. 完成度**のトレードオフも存在します。総括すれば、単純Q&AではGPT-5.1 Instantが群を抜いて速く、複雑推論ではK2やMiniMax-M2が工夫により高いスループットを維持、ERNIEは画像解析込みでも実用的な範囲にある、と言えます。

- **対応言語**: 4モデルは各々の出自から多言語サポートにも特徴があります。**MiniMax-M2**と**Kimi K2**はいずれも中国発のモデルであり、**中国語**での性能が極めて高いと考えられます。実際MiniMax-M2は中

国語から他言語への翻訳で文化的文脈を保持するなど、西洋系モデルにない強みを発揮したとの指摘があります[24]。K2もトレーニングに中国語データを多く含むはずですが、ベンチマークは主に英語で行われています。つまり**英語でトップレベル**を目指す一方、中国語でも一流という二刀流モデルです。**GPT-5.1**はOpenAIのこれまでのモデル同様、英語を筆頭に多言語で優れた能力を持ちます。GPT-4では日本語を含む主要言語でほぼ同等性能でしたし、GPT-5でもそれが受け継がれているでしょう。正式評価は未公表ですが、OpenAIは各国語ユーザからのフィードバックも重視しており、GPT-5.1でも指示への忠実性や文体調整が他言語でも機能するようチューニングされていると思われます。**ERNIE-4.5-VL**はBaidu製ということで**中国語への最適化**が推察されますが、国際ベンチマークである「VLMs Are Blind」や「ChartQA」も攻略していることから、**英語においても一流**であることがわかります[70]。彼らの狙う企業用途（例えば中国国内外の工場データ解析など）では、中国語と英語の両方が飛び交う環境も多いため、ERNIEは両言語対応を重視しているでしょう。全般に、このレベルのLLMになると**トレーニングデータが多言語にまたがる**ため、明示的に「○○語特化版」でない限り複数言語をサポートできます。ただし、専門領域になると各言語で知識量や表現が異なるため注意が必要です。例えば医学や法律の知識は英語情報が豊富ですが、日本語は相対的に少ない、といったケースがあります。そうした領域知識については未だ英語が有利なモデルが多いと考えられます。総じて、**英語**では4モデルともほぼ制約なく利用でき、**中国語**では中国勢モデルがネイティブ性能を発揮、**日本語**を含む他言語ではOpenAIモデルと中国勢モデルが伯仲しつつ、微妙なニュアンスでは訓練データの差が出る可能性がある、といった状況でしょう。

- **API提供状況とコスト**: **GPT-5.1**はOpenAIの商用APIおよびChatGPTサービスで利用可能で、基本料金はGPT-5と同程度と推測されます（執筆時点でOpenAIは細かな料金テーブルを公表していませんが、GPT-4が入力1kトークン0.03ドル、GPT-5がその延長とするとGPT-5.1も近い水準でしょう）。ChatGPT Plusユーザは追加料金なしでGPT-5.1を利用できます[103]。**MiniMax-M2**は公開当初API無料キャンペーンを行い、その後の標準価格を**入力100万トークン0.30ドル・出力100万トークン1.20ドル**としています[7]。この価格はClaude Sonnet 4.5の8%（12.5分の1）と謳われており[7]、非常に安価です。安価ではありますがモデルが冗長に喋りがちなため、長い応答だと実際の課金トークン量が増えてコストが嵩むという指摘もあります[26]。**Kimi K2 Thinking**はMoonshot社のプラットフォームやOpenRouter経由で利用でき、API価格は**入力100万トークン0.60ドル・出力100万トークン2.50ドル**です[39]。MiniMax-M2よりは高いものの、それでもGPT系モデルの10分の1以下、Claude 4.5 Sonnetのおよそ5分の1程度と見られます[7][39]。K2はオープンウェイトでもあるため、十分な計算環境があれば自前運用も可能です。Moonshot社は緩い条件下での商用利用も許可しており（巨大サービスでの表示義務のみ）[35]、企業も組み込みやすいモデルと言えます。**ERNIE-4.5-VL**は完全オープンでApache 2.0ライセンスなので、**API利用料という概念はありません**。誰でもHugging Face等からモデルを落として使えます。ただ実際に動かすには高性能GPUを要するため、Baidu自身もクラウドサービス（例: ERNIE BotやBaidu AI Cloud）の中で提供を始める可能性があります。その場合でも同社が「コスト効率」を前面に出していることから、手頃な価格で使えることが期待されます[125]。なお、オープンモデル全般について言えることですが、サードパーティの推論サービス（例: Together AIやOpenRouterなど）がモデルをホストし独自価格で提供するケースも増えています。ユーザは自前実行か他社経由かを選べ、価格・速度・信頼性のトレードオフを考慮できます。2025年末現在、**最安で最先端モデルを使おうと思えばMiniMax-M2をセルフホスト**するのが一つの解となるでしょう（モデルが無料で提供されているため[126]、電気代とハード代のみで済む）。一方、**手軽さと総合力ではOpenAI GPT-5.1**が依然優位で、安定したAPIとエコシステムが整っています。K2とERNIEはその中間で、強力だがセルフホストのハードルが高め、または第三者サービス経由になるケースが多いという状況です。

2025年11月時点の総括と技術動向 – **生成AIの最先端動向**として注目すべきは、オープンソースモデルの台頭と、それに対抗する形での専有モデルの改良です。2025年11月現在、OpenAIのGPT-5.1やAnthropic Claude 4.5など閉源モデルは依然リーダー的存在ですが、その**性能差は驚くほど縮まっています**[47]。MiniMax-M2やKimi K2、ERNIE-4.5の登場は、「もはや最高性能はオープンコミュニティから生まれ得る」ことを示しました。実際、K2 Thinkingは推論・コーディングでGPT-5に匹敵し、一部では上回る結果を叩き出しました[127]。MiniMax-M2は価格性能比でGPT-5級を実現し、たった数枚のGPUで動く省リソース設計でAGI開発のコ

スト障壁を下げました [6] [8]。ERNIE-4.5-VLはマルチモーダル分野でGoogleのGeminiすら凌駕するタスクがあることを示し [70]、視覚情報を含むAI応用に新たな可能性を拓いています。こうしたオープンモデルの快進撃は、AI研究の地政学的多極化とも相まって、AI開発のイノベーションを加速しています。中国の複数の研究所（Moonshot, MiniMax, Baiduなど）が次々とSOTAモデルを短期間で出してくる様子は、Nathan Lambert氏が「中国の研究所はわずか6ヶ月でオープンフロンティア性能に追いついた」と評するほどで [128]、今後このリストにZhipuAIやAnt Groupなど他の企業も加わり得ると見られています [129]。

一方、**OpenAIやAnthropicといった米国勢**も手をこまねいているわけではありません。OpenAIはGPT-5.1で**ユーザビリティと効率**に舵を切り、より「使って快い」AIへの進化を図りました [81]。またChatGPTを**ユーザが自在にパーソナライズ**できるようにしたことは、単に性能数値では表せない価値を提供しています [82]。AnthropicもClaude 4.5で100k超コンテキストと安定した長文応答を武器に差別化を続けています。**専有モデル**は引き続き閉源ゆえの豊富な資金とフィードバックデータから磨き込まれ、セキュリティ・信頼性・サポート面で企業に選ばれる強みがあります。例えばOpenAIはGPT-5.1リリースと同時にグループチャット機能のパイロットやエージェント安全検証用の模擬システム公開など、エコシステム拡充も進めています [130] [131]。

全体として、**2025年末の生成AI**は「モデルの巨大化＝性能向上」という単純路線から、「より巧みに考え、幅広い入力に対応し、効率良く動くAI」への路線へとシフトしています。コンテキスト拡大、ツール統合、モーダル融合、効率最適化（MoE・量子化）といった多方面のアプローチが組み合わさり、各モデルが独自の進化を遂げています。研究者・技術者にとって、これら最先端モデルの**性能特性と限界**を正確に把握することが重要です。例えば、あるモデルは計算資源が限られる環境で威力を発揮し、別のモデルは特定ドメインのデータ解析で比類なき強さを示す、といった具合に、適材適所でモデルを選ぶ必要が出てきています。また、複数モデルの組み合わせ（例えばGPT-5.1をフロントに、専門タスクではMiniMaxやERNIEを裏で使う）も視野に入るでしょう。2025年11月時点での最先端モデル群を総括すれば、**「オープン vs クローズドの競争が生み出すイノベーション」**が最大の動向と言えます [47] [48]。今後もOpenAIを含む主要プレイヤーとオープンコミュニティが互いに切磋琢磨し、**より高性能で安全かつ用途特化も可能な汎用AI**に近づいていくことが期待されます。

参考文献・情報源:

- MiniMax社「MiniMax M2 & Agent: Ingenious in Simplicity」公式発表 (2025年10月27日) [7] [8]
- Cogni Down Under “MiniMax M2 Beats Gemini at 8% of GPT-5’s Price” (Medium, 2025年11月6日) [6] [25]
- Times of AI “Chinese Open Source AI Model MiniMax-M2 Just Dropped” (2025年10月27日) [12] [132]
- Nathan Lambert “5 Thoughts on Kimi K2 Thinking” (Interconnects, 2025年11月6日) [29] [31]
- VentureBeat (Carl Franzen) “Moonshot’s Kimi K2 Thinking emerges as leading open source AI” (2025年11月6日) [43] [41]
- Towards AI Newsletter #178 (Louie Peters) “Kimi K2 Thinking Steals the Open-Source Crown” (2025年11月13日) [133] [134]
- Baidu ERNIEチーム Hugging Faceモデルカード “Introducing ERNIE-4.5-VL-28B-A3B-Thinking” (2025年11月11日) [57] [135]
- ArtificialIntelligence-News “Baidu ERNIE multimodal AI beats GPT and Gemini in benchmarks” (2025年11月12日) [70] [118]
- OpenAI公式ブログ “GPT-5.1: A smarter, more conversational ChatGPT” (2025年11月12日) [79] [81]
- OpenAI公式ブログ “Introducing GPT-5.1 for developers” (2025年11月13日) [85] [92]
- OpenAI “GPT-5.1 for developers – Model evaluations (Appendix)” (2025年11月13日) [96] [45]

[41] [6]

1 2 3 4 5 6 10 11 23 24 25 26 27 28 106 115 MiniMax M2 Beats Gemini at 8% of GPT-5's Price, But There's a Catch | by Cogni Down Under | Nov, 2025 | Medium
https://medium.com/@cognidownunder/minimax-m2-beats-gemini-at-8-of-gpt-5s-price-but-there-s-a-catch-a438f618bc75

7 8 13 20 21 22 MiniMax M2
https://www.minimax.io/news/minimax-m2

9 107 123 124 MiniMax M2 vs GPT-4o vs Claude 3.5 Benchmark 2025 - Skywork ai
https://skywork.ai/blog/llm/minimax-m2-vs-gpt-4o-vs-claude-3-5-benchmark-2025/

12 126 132 MiniMax-M2 Makes Debut, Beats OpenAI, Anthropic Models
https://www.timesofai.com/news/minimax-m2-ai-launched/

14 MiniMax-M1 - a MiniMaxAI Collection : r/LocalLLaMA - Reddit
https://www.reddit.com/r/LocalLLaMA/comments/1lcuglb/minimaxm1_a_minimaxai_collection/

15 16 GitHub - MiniMax-AI/MiniMax-M1: MiniMax-M1, the world's first open-weight, large-scale hybrid-attention reasoning model.
https://github.com/MiniMax-AI/MiniMax-M1

17 18 19 MiniMax M2 vs M1 2025 Key Upgrades & Why They Matter - Skywork ai
https://skywork.ai/blog/llm/minimax-m2-vs-m1-2025-key-upgrades-why-they-matter/

29 30 31 32 33 34 114 128 129 5 Thoughts on Kimi K2 Thinking - by Nathan Lambert
https://www.interconnects.ai/p/kimi-k2-thinking-what-it-means

35 36 41 42 43 46 47 48 49 50 51 52 53 54 55 108 127 Moonshot's Kimi K2 Thinking emerges as leading open source AI, outperforming GPT-5, Claude Sonnet 4.5 on key benchmarks | VentureBeat
https://venturebeat.com/ai/moonshots-kimi-k2-thinking-emerges-as-leading-open-source-ai-outperforming

37 38 39 40 44 56 113 133 134 TAI #178: Kimi K2 Thinking Steals the Open-Source Crown With a New Agentic Contender | Towards AI
https://towardsai.net/p/machine-learning/tai-178-kimi-k2-thinking-steals-the-open-source-crown-with-a-new-agentic-contender

45 80 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 104 130 131 Introducing GPT-5.1 for developers | OpenAI
https://openai.com/index/gpt-5-1-for-developers/

57 58 60 61 63 64 65 66 67 117 135 baidu/ERNIE-4.5-VL-28B-A3B-Thinking · Hugging Face
https://huggingface.co/baidu/ERNIE-4.5-VL-28B-A3B-Thinking

59 ERNIE-4.5-VL-28B-A3B-Thinking: A Breakthrough in Multimodal AI
https://ernie.baidu.com/blog/posts/ernie-4.5-vl-28b-a3b-thinking/

62 68 69 70 71 72 73 74 75 76 77 109 110 118 119 120 122 125 Baidu ERNIE multimodal AI beats GPT and Gemini in benchmarks
https://www.artificialintelligence-news.com/news/baidu-ernie-multimodal-ai-gpt-and-gemini-benchmarks/

78 Baidu Open-Sources ERNIE-4.5-VL-28B-A3B-Thinking: A Leap in Efficient Visual AI : r/aicuriosity
https://www.reddit.com/r/aicuriosity/comments/1oufjvu/baidu_opensources_ernie45vl28ba3bthinking_a_leap/

79 81 82 103 111 112 GPT-5.1: A smarter, more conversational ChatGPT | OpenAI
https://openai.com/index/gpt-5-1/

83 OpenAI walks a tricky tightrope with GPT-5.1's eight new personalities
https://arstechnica.com/ai/2025/11/openai-walks-a-tricky-tightrope-with-gpt-5-1s-eight-new-personalities/

84 OpenAI says the brand-new GPT-5.1 is 'warmer' and ... - The Verge
https://www.theverge.com/news/802653/openai-gpt-5-1-upgrade-personality-presets

105 OpenAI reboots ChatGPT experience with GPT-5.1 after mixed ...
https://venturebeat.com/ai/openai-reboots-chatgpt-experience-with-gpt-5-1-after-mixed-reviews-of-gpt-5

116 Ernie 4.5-VL-Thinking : Best Open-Sourced Multimodal LLM - Medium
https://medium.com/data-science-in-your-pocket/ernie-4-5-vl-thinking-best-open-sourced-multimodal-llm-21daabfe5f0d

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