2026年中国次世代LLMの技術的・戦略的分析：自律型エージェントと超長文脈処理の最前線

Gemini Opus 3.1

エグゼクティブ・サマリーと市場のコンテキスト

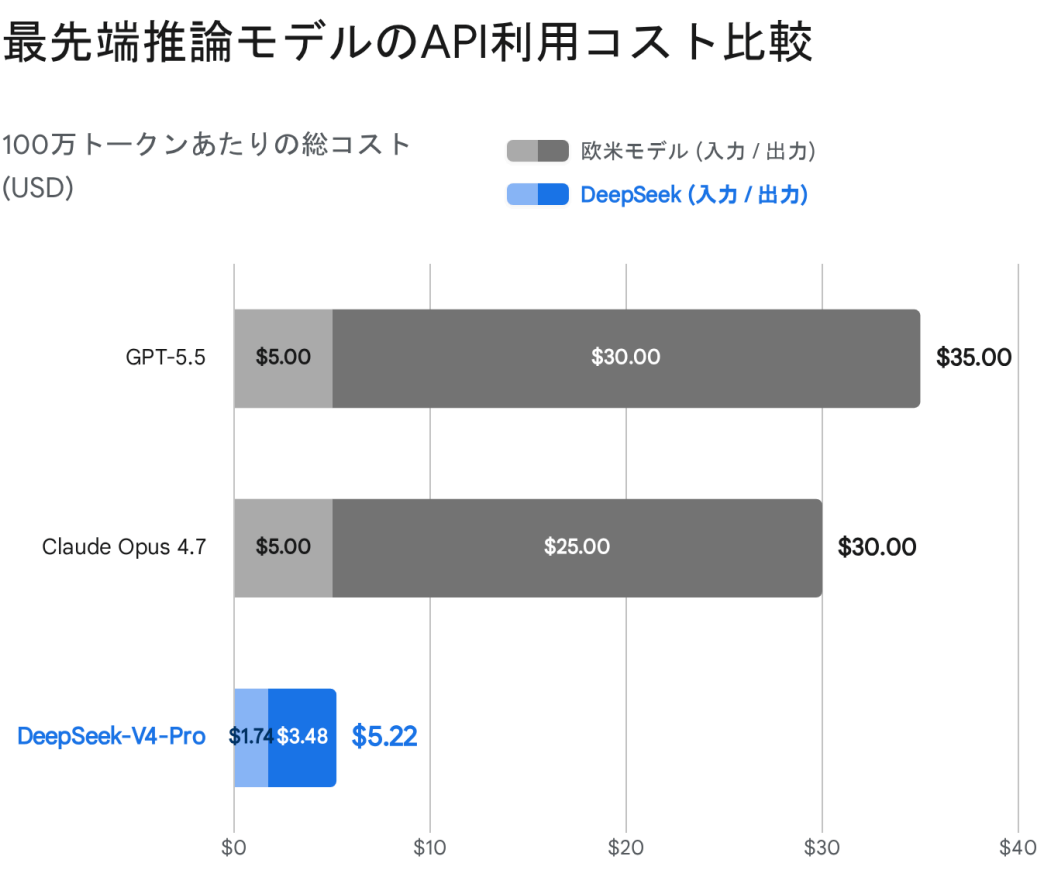
2026年第2四半期現在、世界の基盤モデル市場および人工知能（AI）エコシステムは、根本的なパラダイムシフトの只中にある。2024年から2025年にかけて業界を牽引してきた「パラメータ規模の単純な拡大（スケーリング則）」と「単発のチャット応答品質」の競争は事実上終結し、技術的焦点は完全に「推論時のコンピュート効率（Inference-time Compute）」および「自律型エージェントの長期オーケストレーション能力」へと移行した。このパラダイムシフトを主導しているのが、独自のアーキテクチャ革新によって欧米の最高峰クローズドモデル（GPT-5.5やClaude 4.7 Opus等）と同等以上の性能を、圧倒的な低コストで実現している中国のAI研究所群である[1]。米国主導の主要ラボ（OpenAI、Anthropic、Google等）が依然として一部のベンチマークで優位性を保つ一方で、その差は急速に縮小しており、特定のコーディングおよび推論タスクにおいては逆転現象が起きている[2]。

本レポートは、直近でリリースされた中国発の最先端大規模言語モデル（LLM）4種——DeepSeek「DeepSeek-V4 Preview」、Moonshot AI「Kimi K2.6」、Alibaba「Qwen3.6-27B」、Xiaomi「MiMo-V2.5-Pro」——の技術構造、ベンチマーク性能、および戦略的意義を極めて詳細に分析するものである。これらのモデルは単なるテキスト生成器の枠を完全に超え、数千のツール呼び出しを伴う自律的なソフトウェアエンジニアリング、複数エージェントによる群知能（Swarm Intelligence）の動的展開、ハードウェアおよびIoTデバイスとのネイティブな統合、そして100万トークン規模の超長文脈（Long-context）のインフラ化を達成している[4]。過去のAIシステムが抱えていた、反復的な実行ループ内で文脈を喪失する「金魚の記憶（Goldfish-brained）」問題は、アーキテクチャレベルの再設計によって克服されつつある[8]。各モデルの深層アーキテクチャ設計と市場投下戦略を解剖することで、2026年以降のエンタープライズAIおよび開発者エコシステムが向かう未来図を浮き彫りにする。

DeepSeek-V4 Preview: 限界費用の破壊と超長文脈インフラストラクチャの確立

前世代のV3のローンチから484日後、DeepSeek AI研究所は「第二のDeepSeekモーメント」と称される「DeepSeek-V4 Preview」をリリースし、業界に多大な衝撃を与えた[1]。DeepSeek-V4の最大の功績は、推論コストの破壊的削減と、100万トークンという長大なコンテキストウィンドウを「単なるスペック上の上限」から「日常的に実用可能なインフラストラクチャ」へと再定義したことである[9]。同シリーズは、総パラメータ数1.6兆（アクティブパラメータ49B）の「DeepSeek-V4-Pro」と、総パラメータ数284B（アクティブパラメータ13B）の「DeepSeek-V4-Flash」という2つの強力なMixture-of-Experts（

MoE）言語モデルで構成されている[4]。双方のモデルは完全にMITライセンスの下でオープンソースとして公開されており、商用利用に対しても開かれた姿勢を維持している[11]。



DeepSeek-V4-Proは、キャッシュミス時においてもClaude Opus 4.7やGPT-5.5と比較して約6分の1から7分の1のコストで稼働し、高度なエージェント・ワークフローの経済的制約を根本から取り払う。

データソース: VentureBeat

アーキテクチャの根源的革新：ハイブリッド・スパースアテンションとストレージ最適化

長文脈LLMの構造的なボトルネックは、単にプロンプトに大量のトークンを詰め込めるかどうかではなく、推論時や複雑なエージェントの軌跡（Trajectory）、あるいはツールの連続使用において、モデルがその膨大な文脈に対して「計算資源およびメモリ的に」注意を向け続けられるかどうかにあった[9]。DeepSeek-V4は、最大文脈長をさらに延ばすことよりも、この100万トークンのウィンドウを現実のコンピュート環境でいかに効率的に運用するかに研究の主眼を置いている[9]。

この課題を克服するため、DeepSeekはトークン単位の圧縮技術と「DeepSeek Sparse Attention（DSA）」を組み合わせたハイブリッド圧縮アテンション（Hybrid Compressed Attention）という斬新な機構を採用した[9]。V4-Proの内部構造は全61層のスタックで構成されており、層0から1は高度に圧縮されたアテンションであるHCA（Heavily Compressed Attention）を実行する[12]。続く層2から60にかけては、CSA（Compressed Sparse Attention）とHCAが交互に配置され、計算負荷を動的に分散させる。さらに、ネットワークの終端に位置するMTP（Multi-Token Prediction）ブロックはスライディングウィンドウ処理のみを実行するように設計されている[12]。

このレイヤー配置に加え、ストレージフォーマットの極限までの最適化がKVキャッシュ要件の激減をもたらした。DeepSeek-V4の両アテンションパスは、KV（Key-Value）エントリの大部分の保存にFP8（8ビット浮動小数点）ストレージを採用し、位置情報を司るRoPE（Rotary Position Embedding）の次元にのみ高精度なBF16を使用している[12]。さらに、CSA内部で稼働するライトニングインデクサー（Lightning Indexer）はFP4という極めて低い精度で実行される[12]。これらのストレージ選択と圧縮比率の相乗効果により、従来の標準的なGrouped Query Attention（8ヘッド、BF16保存）のアーキテクチャと比較して、DeepSeek-V4はKVキャッシュのサイズをわずか約2%にまで圧縮することに成功した[12]。

結果として、100万トークンの文脈設定において、V4-Proは前世代のDeepSeek-V3.2と比較して単一トークン推論のFLOPsを27%、KVキャッシュを10%に削減した[4]。より軽量なV4-Flashに至っては、FLOPsを10%、KVキャッシュを7%にまで削減しており、極めて長いコンテキストを扱う展開を飛躍的に容易にしている[12]。

加えて、ネットワーク全体の学習安定性と表現力を高めるための機構的工夫も施されている。従来の残差接続（Residual Connections）を完全に置き換える形で「Manifold-Constrained Hyper-Connections（mHC）」が導入され、モデルの表現力を一切損なうことなく、深い層にわたる信号伝播の安定性が強化された[4]。また、学習の収束速度を加速させ、巨大なパラメータ空間でのトレーニング安定性を担保するためにMuonオプティマイザが採用されている[4]。ポストトレーニングの段階においても、従来の単一ポリシーによる強化学習（Unified-policy RL）から脱却し、独立して訓練されたドメイン特化型専門家モデル（Domain Specialists）群からのオンポリシー蒸留（On-policy Distillation）へとレシピが変更されている[9]。

推論モードの多層化とベンチマークが示す知性

DeepSeek-V4シリーズは、ユーザーのタスクの性質に応じて計算資源の投入量を調整できるよう、「Non-think」「Think High」「Think Max」という3段階の推論エフォートモードをネイティブにサポートしている[4]。

第一の「Non-think」モードは、直感的かつ高速な応答を提供するものであり、リスクの低い日常的なルーチンワークや単純な対話に適している[4]。第二の「Think High」モードは、意識的かつ論理的な分析を行うためのモードであり、応答速度は低下するものの、複雑な問題解決や計画立案において高い精度を発揮する。このモードでは、内部的な思考プロセスが<think> thinking </think> summaryというフォーマットで出力される[4]。第三の「Think Max」モードは、モデルの推論能力の境界を極限まで押し広げるためのモードであり、特別なシステムプロンプトを必要とし、実行には少なくとも384Kトークン分のコンテキストウィンドウの余白が要求される[4]。これらの推論モードを最大限に活用するため

の推奨サンプリングパラメータは、全モード共通でtemperature=1.0、top_p=1.0と指定されている[12]。

これら高度な推論モードを活用したベンチマーク結果は、オープンソースモデルの新たな到達点を示している。DeepSeek-V4-Pro-Maxは、Artificial Analysis Intelligence Indexにおいてスコア52を獲得し、オープンウェイトの推論モデルとしてはKimi K2.6に次ぐ第2位に位置づけられている（前世代のV3.2のスコア42から10ポイントの飛躍的な向上である）[13]。博士課程の専門家レベルの知識を問う超高難度ベンチマーク「GPQA Diamond」において、V4-Pro-Maxは90.1%の正答率を記録し、Claude Opus 4.7（94.2%）やGPT-5.5（93.6%）といったトップクラスのクローズドモデルに肉薄している[14]。また、数学的能力を測る「AIME 2026」や「HMMT 2026 Feb」においても、それぞれ非常に高い問題解決能力を実証している（HMMT 2026 Febで95.2%の正答率）[5]。

開発現場における実用性も高く評価されている。DeepSeekの内部研究開発コーディングベンチマーク（PyTorch、CUDA、Rust、C++にまたがる30の厳選されたタスク）において、V4-Pro-Maxは67%のパスレートを達成し、Claude 3.5 Sonnet（47%）を大幅に上回り、Opus 4.5（70%）に迫る結果を出した[12]。さらに、V4-Proを日常的に使用している85名のDeepSeek開発者を対象とした調査では、52%が「現在のメインのコーディングモデルを置き換える準備ができている」と回答し、39%が「そうする傾向にある」と答えており、実務レベルでの信頼性が証明されている[12]。長文脈の検索能力を示すMRCR 8-needleテストにおいても、V4-Pro-Maxは256Kトークンまで精度0.82以上を維持し、100万トークン時点でも0.59の精度を保持している[12]。

エージェントループにおける限界費用の破壊

DeepSeek-V4が市場に与える最も直接的なインパクトは、API価格の劇的な破壊を通じた「エージェント経済学」の再定義である。DeepSeek-V4-ProのAPI利用料金は、キャッシュミス時の100万入力トークンあたり1.74ドル、100万出力トークンあたり3.48ドルに設定されている[1]。単純な100万入力・100万出力の組み合わせで比較した場合、総コストは5.22ドルとなる。さらに、キャッシュヒット時には入力価格が0.145ドルにまで急減し、同条件での混合コストはわずか3.625ドルに収まる[1]。

この価格体系は、GPT-5.5（入力5ドル／出力30ドル、合計35ドル）やClaude Opus 4.7（入力5ドル／出力25ドル、合計30ドル）といった米国の最先端モデルと比較して、約6分の1から7分の1のコストで同等の推論能力にアクセスできることを意味する[1]。また、オープンソース開発者コミュニティで行われた7つのモデル（GLM-5.1、Kimi K2.6、MiMo-V2.5、MiniMax M2.7、Qwen3.6 Plus、DeepSeek V4 Pro、DeepSeek V4 Flash）を対象としたOpenCode Goを用いたコマンドライン（CLI）ベンチマークにおいては、非常に興味深い実践的な結果が示された[16]。

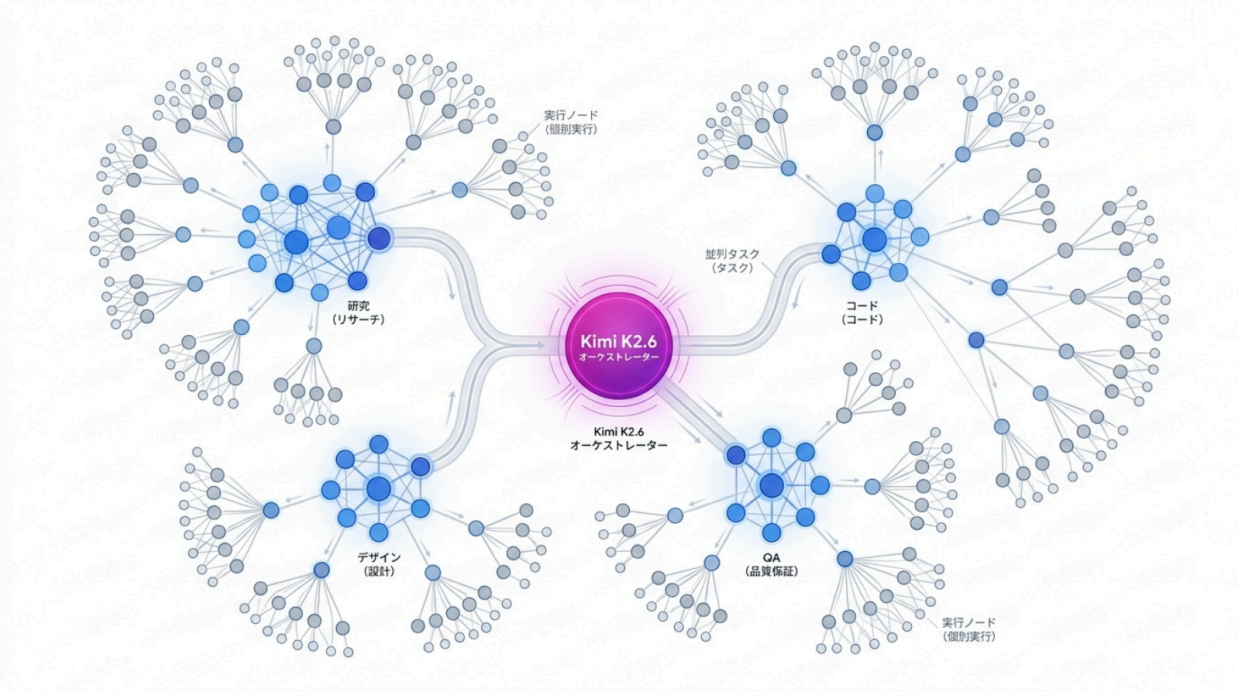
このテストでは、知能、推論能力、価格、文脈長、速度などの要素を総合的にブレンドして評価が行われた。その結果、最も高度な推論能力を持つ「DeepSeek-V4-Pro」を抑え、より軽量な「DeepSeek-V4-Flash」が第1位にランクインしたのである[16]。これは、何時間にもわたってコードの編集・テスト・修正を何千回も繰り返す自律型エージェントのループ処理においては、知能のわずかな優位性よりも、コストとレイテンシの低さ、そして100万トークンの文脈を維持できるインフラ性能の累積的なメリットが上回ることを示している[16]。V4-Flashは、この「十分な推論能力と圧倒的な低コスト・高速性」のスイートスポットを突き、大量のトランザクションを発生させるエージェントワークフローのデフォルトエンジンとしての地位を確立しつつある。

Moonshot Kimi K2.6: スウォーム知能とマルチモーダル自律実行の極致

Moonshot AIによって開発された「Kimi K2.6」は、総パラメータ数1兆（トークンあたり32Bがアクティブ）の強大なMixture-of-Expertsアーキテクチャに、400MパラメータのMoonViTビジョンエンコーダを統合した、オープンソースのネイティブ・マルチモーダル・エージェントモデルである[17]。全61層、384の専門家ネットワーク（トークンごとに8つが選択され、1つが共有される）で構成されており、262,144トークンのコンテキストウィンドウを備え、長時間のコーディング、コーディング主導の設計、プロアクティブな自律実行において驚異的な能力を発揮する[17]。

Kimi K2.6は、DeepSeekが推論効率に極振りしたのとは対照的に、「エージェントの自律的オーケストレーション」と「マルチモーダルな知覚」の統合に戦略の主眼を置いている。Artificial Analysis Intelligence Indexにおいてはスコア54を記録し、オープンウェイトモデルとして現時点でトップの座を確保している[13]。

スウォーム・アーキテクチャによるタスクの並列分解と自律実行



Kimi K2.6は複雑なタスクを動的に分解し、最大300のサブエージェントを水平展開させることで、4,000の調整されたステップを並行して実行する。

スウォーム・オーケストレーション：「スケールアップ」から「スケールアウト」への設計思想の転換

Kimi K2.6を他モデルから明確に差別化する最大の特徴は、「Agent Swarm（エージェント群）」アーキテクチャによる水平方向への拡張性である。Moonshot AIは、モデル単体の処理能力を向上させる「スケールアップ」のみに頼るのではなく、作業を分散させる「スケールアウト」の哲学を深く実装している[5]。

このシステムは、与えられた複雑なタスクを動的に細かいサブタスクに分解し、モデル自身が生成した専門的なエージェント群にそれらを割り当てて同時並行で実行させる。アーキテクチャの進化により、K2.6は前世代のK2.5（最大100エージェント、1,500ステップ）から大幅にスケーラビリティを向上させ、最大300のサブエージェントが4,000の調整されたステップを同時に実行できるようになった[5]。この構成的な知能（Compositional Intelligence）の連携により、広範なウェブ検索と深い学術調査の組み合わせや、大量の文書分析から複数フォーマットのコンテンツ生成まで、レイテンシを削減しつつ出力品質を劇的に高めることが可能となっている[5]。

スウォーム・オーケストレーションの実社会での適用例は多岐にわたる。例えば、ある候補者の履歴書（CV）をカリフォルニア州の100の関連する求人ポジションと照合するために100のサブエージェントを生成し、それぞれの役割に合わせたカスタマイズされた履歴書を同時生成するタスクが報告されている[5]。また、100のグローバルな金融資産に対して定量的な戦略を実行し、世界トップクラスのコンサルティングファームに匹敵するプレゼンテーション資料を一括で作成するプロセスや、Googleマップ上で30の小売店舗を特定し、それぞれの店舗に最適化されたコンバージョン率の高いランディングページを自律的に構築するワークフローなど、従来の人海戦術を無力化する規模の実行力が証明されている[5]。

さらに、K2.6の特異な能力として、PDF、スプレッドシート、Wordドキュメントなどの高品質なファイルから、その構造的・文体的なDNAを抽出し、再利用可能な「スキル（Skills）」としてシステム内に蓄積する機能も有しており、エージェント群のタスク遂行能力を継続的に進化させることができる[5]。

プロアクティブな長期コーディング能力とシステム改修

Kimi K2.6は、フロントエンド開発、DevOps、パフォーマンス最適化など、極めて長期的な視座が求められるソフトウェアエンジニアリングタスク（Long-horizon coding）において、Rust、Go、Python、さらにはZigなどの多様なプログラミング言語間で信頼性の高い汎化能力を示している[5]。

特筆すべき実証例として、Zig言語を用いてMac環境上にQwen3.5-0.8Bモデルのローカル推論システムをデプロイしたパフォーマンス最適化のケースがある。K2.6は12時間かけて14回のイテレーションを行い、その過程で4,000回のツール呼び出しを実行した。結果として、スループットを初期の約15 tokens/secから約193 tokens/secへと飛躍的に向上させ、既存のオープンソースツール（LM Studioなど）を20%上回る性能を叩き出した[5]。また、8年前に記述された「exchange-core」と呼ばれる金融マッチングエンジンの全面的な改修タスクにおいては、13時間に及ぶ自律セッションを通じて4,000行以上のコードを変更し、コアスレッドのトポロジを再構成することで、中程度のスループットを185%向上させるという離れ業をやってのけている[5]。内部評価テストでは、コード生成精度の12%向上に加え、ツール呼び出しの成功率が96.60%に達することが確認されている[5]。

K2.6の自律性は単発のタスクにとどまらない。同モデルは「OpenClaw」や「Hermes」といったプロアクティブなエージェント・フレームワークをバックエンドで駆動し、人間の監視を一切必要とせず、24時間365日バックグラウンドで動作し続ける能力を備えている[5]。実際に、K2.6を組み込んだエージェントが5日間にわたり自律的に稼働し、システムの監視、インシデントへの対応、アラートの発報から問題解決に至る全サイクルを独力で管理した事例が報告されている[5]。さらに、この研究プレビューの一環である「Claw Groups」機能では、複数の人間と様々なデバイスやモデルからのエージェントが共有のオペレーションスペースで真の共同作業者として活動し、K2.6が適応型のコーディネーターとして機能して、停止したエージェントからタスクを再割り当てする高度な管理能力を見せている[5]。

マルチモーダル推論と戦略的「Modified MIT License」

視覚入力を用いた「コーディング主導の設計（Coding-Driven Design）」もK2.6の真骨頂である。単純なテキストプロンプトや画像入力から、完全にインタラクティブなフロントエンドのインターフェースや、認証・データベース操作を含むフルスタックのワークフローを生成する[5]。画像・動画生成ツールと連携し、ウェブサイト用の審美的なヒーローセクションやスクロール連動型アニメーションなど、視覚的に一貫性のあるアセットを構築する能力も持つ[5]。Pythonツールを活用した推論モードをオンにした際の視覚ベンチマークスコアは圧倒的であり、MathVisionで93.2、V*で96.9、MMMU-Proで80.1を記録している[5]。また、HLE（Humanity's Last Exam）のツール利用フルセットにおいて54.0を記録し、GPT-5.4（52.1）やClaude Opus 4.6（53.0）を上回る推論力を示している[5]。

Moonshot AIは、これらの強大な機能を社会実装するにあたり、ライセンス戦略においても独自の工夫を凝らしている。Kimi K2.6は「Modified MIT License」の下で提供されており、大部分のユースケースにおいて無料かつ自由に利用、改変、再配布が可能である[21]。しかし、極めて重要な変更点として、「月間アクティブユーザー数が1億人を超える、または月間収益が2000万米ドルを超える商用製品・サービス」にこのソフトウェア（またはその派生物）を組み込んで使用する場合、ユーザーインターフェース上に「Kimi K2」の文字を明確に表示しなければならないというアトリビューション条項が追加されている[21]。この条項は、資本力に勝る巨大テック企業による無断でのフリーライドを牽制しつつ、AIエコシステムの最上位層に自社ブランドを強制的に浸透させる高度な知財・マーケティング戦略と評価できる。なお、この制限条項は合成データ（Synthetic data）の生成や、合成データを用いて訓練された派生モデルには適用されないことが確認されており、オープンソースコミュニティにおけるモデルの継続的な進化を妨げないように配慮されている[22]。

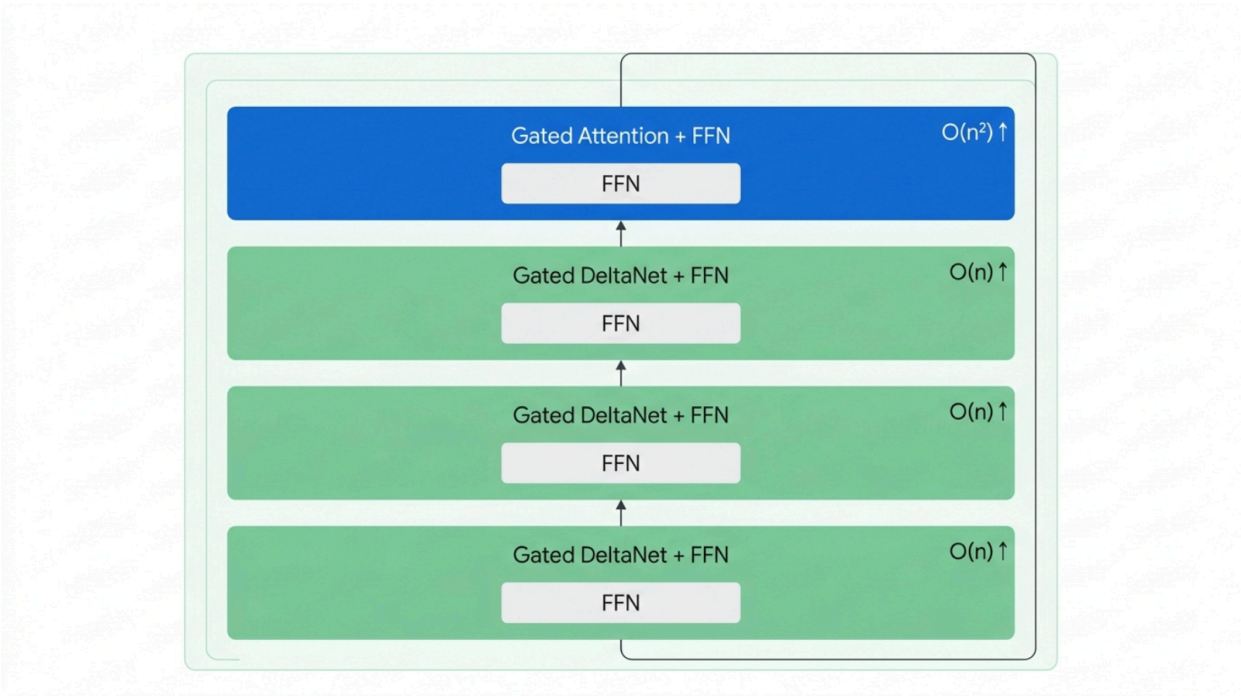
Alibaba Qwen3.6-27B: Gated DeltaNetによるアテンションの線形化と推論プロセスの保存

Alibaba CloudのQwenチームがリリースした「Qwen3.6」シリーズは、前世代のQwen3.5のブレイクスルーを基盤としつつ、実際の開発現場での安定性と実用性に極めて高い優先度を置いたアップデートである[23]。その中でも特に注目されるのが、270億（27B）という比較的小規模なパラメータ数を持つDense（密）アーキテクチャでありながら、397Bクラスの巨大MoEモデル（Qwen3.5-397B-A17B）をエージェント・コーディング能力で凌駕する「Qwen3.6-27B」である[7]。同シリーズには、35Bの総パラメータ（アクティブ3B、256エキスパート）を持つハイブリッド推論モデル「Qwen3.6-35B-A3B」もラインナップされており、22GBのRAM環境で稼働可能である[8]。

ハイブリッド・アテンション：線形（Linear）とフルSoftmaxの精緻な融合

Qwen3.6-27Bのアーキテクチャの核心は、Transformerの構造的欠陥とも言える「コンテキスト長の二次関数的な計算量増加」を根本から解決するための「Gated DeltaNet」の導入である[7]。従来の標準的なSoftmax Attentionは、入力シーケンスが長くなるにつれて計算量とメモリ消費が$O(n^2)$で爆発するため、超長文脈の処理において深刻なボトルネックとなっていた。Qwen3.6は、リカレントニューラルネットワーク（RNN）やMambaアーキテクチャから着想を得た線形アテンション（Linear Attention）機構をハイブリッドで導入することで、この複雑性を$O(n)$へと線形化している[7]。

Qwen3.6 ハイブリッド・アテンションのブロック構造



Qwen3.6は、計算効率の高い線形アテンション（Gated DeltaNet）と高精度なフルアテンション（Gated Attention）を3:1の割合でハイブリッド配置し、コンテキスト長の二次関数的な計算爆発を回避している。

詳細な隠れ層のレイアウトは、全64層のネットワーク内で「16 × (3 × (Gated DeltaNet → FFN) → 1 × (Gated Attention → FFN))」という特殊な比率を持つマクロブロックの繰り返しで構成されている[7]。すなわち、4つのサブレイヤーのうち3つを計算量が$O(n)$で済む高速かつ状態保持に優れた「Gated DeltaNet」に置き換え、残りの1つに高い推論精度を担保する従来の「Gated Attention（フルSoftmax）」を配置する3:1の比率を採用している[7]。これは妥協の産物ではなく、長大なシーケンスを安価に処理する能力と、複雑な論理展開を精密に解釈する能力を両立させるための意図的な設計である[8]。

Gated DeltaNet層には、いつ過去の情報を更新し、いつ保持するかをネットワーク自身に学習させるゲーティング・メカニズム（LSTMのゲートに概念的に類似）が組み込まれている[7]。このサブレイヤーでは、Value（V）次元に48の線形アテンションヘッド、QueryとKey（QK）次元に16のヘッドが割り当てられ、ヘッド次元は128に設定されている[7]。一方、精緻な解釈を担うGated Attention層では、Query（Q）に24ヘッド、KVに4ヘッドが割り当てられ、ヘッド次元は256とより大きく確保されている[27]。位置エンコーディング（RoPE）の次元は64であり、Feed Forward Network（FFN）の中間次元は17,408に達する[27]。また、学習の安定化のためにZero-Centered RMSNormが採用されている[28]。

このハイブリッド構造に加え、推論時のサービングにおいて投機的デコーディング（Speculative Decoding）を可能にし、トークン生成速度を大幅に引き上げるMulti-Token Prediction（MTP）技術も導入されている[7]。これにより、Qwen3.6-27Bは推論時のスループットを高く維持しながら、ネイティブで262,144トークン、YaRN技術の適用により最大1,010,000トークンという超長文脈を極めて効率的に処理することが可能となった[7]。

Thinking Preservation（推論プロセスの保存）による自律ループの安定化

開発者からの直接的なフィードバックを受けて実装されたQwen3.6の機能的ブレイクスルーが、「Thinking Preservation（思考の保存）」オプションである[23]。従来のLLMを活用したエージェント開発においては、段階的なリファクタリングやマルチステップのデバッグといった反復的なプロセスの中で、モデルはターンのたびに過去のコンテキストからゼロベースで推論（思考）を構築し直す必要があった。この無駄な再計算は膨大なコンテキストオーバーヘッドを生み出すだけでなく、エージェントが過去の文脈を突然失い、同じエラーを繰り返す「ループによる失敗（loop-and-repeat failures）」の主要な原因となっていた[8]。

Thinking Preservationを有効にすると、過去の対話履歴から「推論コンテキスト（推論の痕跡）」がそのまま保持され、次ターンの推論の基盤として引き継がれる[23]。これにより、モデルの記憶喪失（Goldfish-brained状態）が防がれ、リポジトリレベルの推論やフロントエンドの複雑なワークフローにおいて、プロンプトエンジニアリングによる回避策を必要とせずに、一貫性のある長期的な計画と反復開発が実現する[8]。このパラメーターはデフォルトではオフに設定されているが、API経由で容易に有効化できる[8]。Qwen3.6は、OpenClaw（旧Moltbot/Clawdbot）やClaude Code、Qwen Codeといった主要なサードパーティ製コーディングアシスタントとシームレスに統合可能であり、ターミナル環境で完全なエージェント・コーディング体験を開発者に提供している[29]。また、201の言語と方言をサポートするグローバルな言語カバレッジも前世代から継承されている[23]。

Xiaomi MiMo-V2.5-Pro: エコシステム統合とマルチモーダル・エージェントの民主化

スマートフォンおよびスマート家電の世界的巨人であるXiaomiが自社開発し、投入した「MiMo-V2.5-Pro」および「MiMo-V2.5」は、APIプラットフォームでの汎用的な提供に加え、自社のOSエコシステム（HyperOS 2.0）や膨大なハードウェア群とのネイティブな融合を前提として設計された、独自の戦略的ポジショニングを持つネイティブ・マルチモーダルモデルである[30]。

MiMo-V2.5シリーズの開発哲学は、テキストだけでなく画像や音声をネイティブに理解し、知覚した

情報に基づいて直接行動を起こす「一つのモデルで全てを理解し、実行する」ことにある[31]。以前はOpenRouter上で「Hunter Alpha」という匿名名義でテスト運用され、一日のAPIコール数チャートでトップに立つなど、公開前からその実力は広く認知されていた[33]。

エッジ・ハードウェアを制御する「miclaw」AIアシスタントの頭脳

MiMo-V2.5シリーズのエンタープライズ領域における最大の強みは、Xiaomiが推進する「Human x Car x Home（人・車・家）」という広大なスマートエコシステムを自律的に制御する次世代AIアシスタント、「miclaw」のバックエンド頭脳として機能する点にある[32]。現在グローバルでクローズドベータテストが進行中のmiclawは、従来の単なる対話型音声インターフェース（Xiao AIなど）の概念を破壊し、システムレベルでの自律的な実行能力（System-level execution）を提供する[32]。

miclawの動作プロセスは、推論と実行のループ（Inference-execution loop）に基づいて構築されている[32]。ユーザーの意図を分析すると、自ら適切なツールとパラメータを選択してタスクを実行し、その結果をレビューして完了するまで反復的に作業を続ける[32]。さらに、Model Context Protocolを採用することで、デバイス内の他の主要なプロセスを妨げることなく既存のAIユーティリティと効率的に連携する[32]。

Mi Homeアプリとの統合により、miclawは家中のスマートデバイス（ルーター、監視カメラ、フロアクリーナーなど）のステータスを読み取り、コンテキストに基づく高度な判断を下す[32]。例えば、ユーザーのカレンダーに「会議」の予定が入っている場合、従来のアシスタントは一律にスマートフォンをサイレントモードにするだけであった。しかしmiclawは、カレンダーのラベルからそれが「重要な顧客との会議」なのか「社内の定例会議」なのかを解釈し、過去の習慣や接続されたデバイスの現在のステータスを総合的に判断して、必要であれば家中の関連デバイス（照明、空調、インターホン等）の動作を連携して調整するレベルの自律制御を実現している[37]。この進化の背景には、過去のMIUI時代からの技術的負債（Legacy SDKs）を徹底的に排除し、安定性とAIネイティブな設計を追求したHyperOSの「Zero-Legacy Architecture」の恩恵がある[34]。

フロンティア級のマルチモーダル知覚と長期開発能力

モデル単体の知能テストにおいても、MiMo-V2.5シリーズは世界の最先端モデル群に肉薄している。フラッグシップであるMiMo-V2.5-Proは1兆を超えるパラメータ（アクティブ42B）を持ち、100万トークンのネイティブコンテキストをサポートする[33]。GPQA Diamondで86.6%、SciCodeで50.2%、IFBenchで79.9%という高いスコアを叩き出し、Artificial Analysis Intelligence Indexでは中国LLMとして2位にランクインしている[33]。

特筆すべきは、MiMo-V2.5（標準モデル）の卓越したマルチモーダル理解能力である[31]。会議の録画、製品デモ、スポーツ映像、ドキュメンタリーなど、数分間に及ぶ長尺の動画素材からのシーントラッキング、時間的推論、および視覚的グラウンディングの能力を測る「Video-MME」ベンチマークにおいて、MiMo-V2.5は87.7を記録した。これは単なるキャプション生成ではなく真の映像理解を示すスコアであり、GoogleのGemini 3 Pro（88.4）と実質的に同等レベルであり、Gemini 3 Flashを大きく引き離している[31]。画像理解においても、CharXiv RQで81.0、MMMU-Proで77.9を記録し、業界トップ水準に迫っている[31]。

エージェントを用いた自律的なコーディング環境においては、さらに驚異的な持続力（Long-horizon

autonomy）を証明している。公式のデモンストレーションにおいて、MiMo-V2.5-Proは単一のプロンプトから自律的に11.5時間稼働し続け、その間に1,868回のツールコールを実行して、マルチトラック・タイムライン、クリップのトリミング、クロスフェード、音声ミキシング、そしてエクスポート・パイプラインを備えた完全なデスクトップ動画編集アプリ（総コード数8,192行、MiMo-V2-TTSによるAIボイスオーバー付き）をゼロから構築することに成功した[6]。また、Rust言語を用いてSysYコンパイラを構築するタスク（233/233テストクリア、672ツールコール、4.3時間）も完遂しており、複雑なソフトウェアの全体構造を破綻なく維持する能力が実証されている[39]。

経済性と開発者へのアプローチも非常に戦略的である。MiMo APIの価格設定は、100万入力トークンあたり0.40ドル、出力2.00ドル（標準モデル）という安価な設定であり、長文脈の100万トークンウィンドウ利用に対する追加のマルチプライヤー課金を廃止している[33]。さらに、クレジットカード登録不要で利用可能な100%無料のAPIティア（mimo-v2.5-pro）を提供し、OpenAI互換のエンドポイントを通じてVS CodeのClineやRoo Codeといった開発エディタ拡張に即座に組み込める体制を整えている[40]。これにより、開発現場におけるClaude 3.5 Sonnet等の既存クローズドモデルのシェアを強力に切り崩す動きを見せている[40]。

各モデルの統合的なベンチマークおよび特性比較

2026年第2四半期における自律型エージェント環境の覇権を争う主要モデル群の性能と特性を、定量的な指標を用いて比較する。これらのデータは、各社がどのようなアーキテクチャ哲学に基づき、どの指標の最大化を狙っているかを明確に示している。

実務遂行およびエージェント自律能力の比較分析

以下の表は、実際の開発リポジトリの問題解決能力を測るSWE-Benchや、ターミナル環境での自律実行力を測るTerminal-Bench 2.0など、近年のAI評価において最も重視される「エージェント型ワークフローの達成度」を比較したものである。

評価指標・ベンチマーク	DeepSeek-V4-Pro (Max)	Kimi K2.6 (Thinking)	Qwen3.6-27B (Plus/Max)	MiMo-V2.5-Pro	欧米主要モデル参考値
SWE-Bench Verified (解決率)	80.6% [14]	80.2% [14]	-	-	Claude 4.6 Opus: 80.8% [14]
SWE-Bench Pro (解決率)	55.4% [14]	58.6% [14]	-	57.2% [39]	GPT-5.4: 57.7% [14]
Terminal-Bench 2.0	67.9% [14]	66.7% [14]	-	43.2% [33]	GPT-5.4: 75.1% [14]

(精度)					Claude 4.6 Opus: 65.4% [14]
GDPval-AA (Eloレーティング)	1554 [13]	1482 [14]	-	-	GPT-5.4: 1674 [14] Claude 4.6 Opus: 1619 [14]
Claw-Eval (一般的なタスク)	-	-	-	63.8 [39]	-

このデータ構造から読み取れる極めて重要なインサイトは、DeepSeek-V4-ProとKimi K2.6が、実世界の複雑なソフトウェアリポジトリ内のバグ修正能力を測る「SWE-Bench Verified」において、いずれも80%の壁を突破し、Anthropicの最高峰であるClaude 4.6 Opus（80.8%）と完全に同格の実力を持っているという事実である。ターミナル操作を伴う「Terminal-Bench 2.0」においては、GPT-5.4が75.1%と頭一つ抜けているものの、中国発のモデル群も60%台後半のスコアで追従しており、オープンソースおよび安価なAPIでここまでの実行精度が担保されている事実は、業界構造を根底から揺るがすものである。

モデルアーキテクチャおよび機能特性のサマリー

特徴・システム構成	**DeepSeek-V4-Pro**	**Moonshot Kimi K2.6**	**Alibaba Qwen3.6-27B**	**Xiaomi MiMo-V2.5-Pro**
総パラメータ規模	1.6兆 (アクティブ49B) [4]	1兆 (アクティブ32B) [17]	27B (Dense アーキテクチャ) [24]	1兆超 (アクティブ42B) [33]
アーキテクチャの核心	ハイブリッド圧縮アテンション (CSA/HCA) [9]	ネイティブ・マルチモーダルMoE + MoonViT [18]	Gated DeltaNet (O(n) 線形アテンション) [7]	視覚・音声エンコーダ統合型 MoE [31]
最大サポート文脈長	100万トークン [11]	262,144トークン [17]	256,000 (YaRNで最大101万) [7]	100万トークン [31]

エージェント運用上の特筆機能	推論モードの動的切替（Max/High/Non-think）、極小KVキャッシュ要件 [4]	スウォーム並列オーケストレーション（最大300エージェント、4000ステップ） [5]	Thinking Preservation（対話履歴越しの推論コンテキスト保存） [23]	miclaw統合によるIoT自律連動、長尺動画の時空間推論理解 [31]

戦略的洞察と2026年以降のエンタープライズAIの展望

本レポートで提示された技術的仕様と各社の動向を俯瞰すると、2026年のAI基盤モデル市場を決定づける幾つかの深層的（二次的・三次的）なメガトレンドが浮き彫りになる。

1. 長期自律性（Long-Horizon Autonomy）の確立とループ経済の支配

これまで、LLMを用いた自律エージェントは、せいぜい数十回のツール呼び出しやステップを経た段階で、文脈の破綻や情報の「忘却」を起こし、タスクの完遂に失敗していた。しかし、XiaomiのMiMo-V2.5-Proが示した「11.5時間・1,868回の呼び出しによる動画エディタの完全構築」[6]や、Moonshot Kimi K2.6による「5日間にわたる無人システム運用」[5]が証明するように、エージェントの「安全稼働限界時間」は数時間から数日単位へと劇的に延長された。

この進化の根底にあるのは、DeepSeek-V4のKVキャッシュ圧縮技術（標準比2%）[12]や、Qwen3.6の線形アテンション（Gated DeltaNet）[7]といった、アーキテクチャレベルでの計算複雑性の排除である。ソフトウェア開発のループ（テスト・編集・コンパイル・エラー修正の絶え間ない繰り返し）においては、モデル呼び出しごとのレイテンシとAPIコストが雪だるま式に累積する。そのため、「極めて賢いが計算が重く高価なモデル」よりも、「十分に賢く、かつ圧倒的に安価で高速なモデル」（例えばDeepSeek V4 Flashなど）の方が、長時間の自律エージェントのバックエンドエンジンとしては実務上の価値が高いという評価基準の完全な逆転現象が起きている[16]。Qwen3.6の「Thinking Preservation」による推論オーバーヘッドの削減も、このループ経済の効率化に直結している[23]。

2. スウォーム知能による「時間と精度のトレードオフ」の突破

Moonshot AIがKimi K2.6で提示した「最大300のサブエージェントによる並列実行（Agent Swarm Orchestration）」は、生成AIの社会実装手法を根本から変容させるパラダイムである[5]。これまでのAI活用は、「より巨大で優れた1つのモデル」に複雑な長文プロンプトを解釈させ、直列的に推論を行わせる（スケールアップ）手法が主流であった。しかし、Kimi K2.6はマクロな指示を動的に分割し、役割に特化させた多数のエージェント群に同時並行で処理させる（スケールアウト）というアプローチをとる。

これにより、例えば「市場調査・競合分析・コード生成・デザイン適用・資料作成」といった本来ならば複数部門にまたがる高度に構造化されたエンタープライズ・ワークフローを、人間の介入なしに一気通貫かつ極めて短時間で生成することが可能となった。これは、単一のエージェントの直列的な処理能力の限界を「群の力（Swarm Intelligence）」で突破するものであり、企業の生産性向上を指数関数的に加速させる可能性を秘めている。

3. ハードウェア／OSレベルへの「知能の溶解」

Xiaomiの「MiMo-V2.5」および「miclaw」AIアシスタントの動向は、LLMがクラウド上のブラウザベースのチャットUIという狭い枠から脱却し、ハードウェアやオペレーティングシステム（HyperOS 2.0等）とネイティブに統合されつつある未来を示唆している[32]。モデルがカメラの映像、マイクの音声、そして家庭や車内にある数百万のIoTデバイスから直接コンテキストをリアルタイムで読み取り、デバイス間で複雑なアクションを自律的にオーケストレーションする段階に入っている[37]。Qwen3.6のZigを用いたエッジデバイス（Mac等）へのローカルデプロイ環境の最適化事例も、クラウドコンピュート依存からの脱却と、エッジ・オンデマイスAIの高度化を裏付ける強力な証拠である[5]。

4. 限界費用の不可逆的な低下とAI開発競争の地政学

DeepSeek-V4-ProがGPT-5.5やClaude 4.7 Opusの約6分の1から7分の1という破壊的なコストで提供されている事実は、米中間のAI開発競争において極めて重要な地政学的および経済的意味を持つ[1]。中国のAI企業は、米国の先進半導体輸出規制等により、自由にアクセスできるコンピュート資源（GPU等）に厳格な制約がある環境下で開発を進めざるを得なかった。

しかし皮肉なことに、この厳しいハードウェア制約が、結果として研究者たちに「アーキテクチャの徹底的な効率化（MoEの高度化、MTPの採用、ハイブリッド線形アテンション、FP8/FP4ストレージの活用等）」を限界まで追求させる強いインセンティブとして働き、結果として「極めて少ない計算量で世界の最高水準の推論を引き出す」独自の技術体系を鍛え上げるに至った[3]。この「効率性のイノベーション」が、膨大な資本を投下して力技でモデルを肥大化させてきた欧米のフロンティアモデルの価格決定権を根本から揺るがし、エージェントAIの民主化を強制的に推進する原動力となっている。

結論

2026年第2四半期に登場した中国発の次世代LLM群（DeepSeek-V4、Kimi K2.6、Qwen3.6、MiMo-V2.5）は、単なるパラメータ規模のインフレーションという力技の競争を脱却し、「斬新なアーキテクチャによる限界費用の劇的な引き下げ」「100万トークン規模の長文脈インフラストラクチャ化」「群知能（Swarm）とエッジIoTを巻き込んだ自律型エージェントの社会実装」という、実用性に直結する三つの新たなフロンティアを同時に切り拓いた。

企業のITリーダーやソフトウェア開発者にとって、推論コストの桁違いの低下とエージェントの長期稼働能力の成熟は、AIの役割を「人間の作業を補助する単発の質問応答ツール」から「24時間365日自律的に稼働し、自己修正を繰り返しながらタスクを完遂するデジタルな労働力」へと完全にシフトさせることを意味する。今後の市場における競争の優位性は、もはや「どのベンダーのモデルを利用するか」という選択次元にはなく、「これらの安価で強力な自律型コンポーネントを、自社のビジネスロジックに適合した高度なオーケストレーション・フレームワーク（スウォーム構築やIoT制御等）にいかに深く組み込み、独自の価値を生むワークフローを構築できるか」という、システムエンジニアリングとインテグレーションの領域へと不可逆的に移行していくであろう。

引用文献

1. DeepSeek-V4 arrives with near state-of-the-art intelligence at 1/6th the cost of Opus 4.7, GPT-5.5 | VentureBeat, 4月 26, 2026にアクセス、

https://venturebeat.com/technology/deepseek-v4-arrives-with-near-state-of-the-art-intelligence-at-1-6th-the-cost-of-opus-4-7-gpt-5-5
2. AI Trends 2026 – LLM Statistics & Industry Insights, 4月 26, 2026にアクセス、https://llm-stats.com/ai-trends
3. DeepSeek is back with V4, slashing agentic AI costs - Techzine Global, 4月 26, 2026にアクセス、https://www.techzine.eu/blogs/analytics/140764/deepseek-is-back-with-v4-slashing-agentic-ai-costs/
4. deepseek-ai/DeepSeek-V4-Pro - Hugging Face, 4月 26, 2026にアクセス、https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro
5. Kimi K2.6 Tech Blog: Advancing Open-Source Coding - Kimi AI, 4月 26, 2026にアクセス、https://www.kimi.com/blog/kimi-k2-6
6. Xiaomi MiMo-V2.5-Pro, 4月 26, 2026にアクセス、https://mimo.xiaomi.com/mimo-v2-5-pro/
7. Alibaba Qwen Team Releases Qwen3.6-27B: A Dense Open-Weight Model Outperforming 397B MoE on Agentic Coding Benchmarks - MarkTechPost, 4月 26, 2026にアクセス、https://www.marktechpost.com/2026/04/22/alibaba-qwen-team-releases-qwen3-6-27b-a-dense-open-weight-model-outperforming-397b-moe-on-agentic-coding-benchmarks/
8. Your AI Agent Is Goldfish-Brained. Qwen3.6–35B-A3B Is the Fix. - Towards AI, 4月 26, 2026にアクセス、https://pub.towardsai.net/your-ai-agent-is-goldfish-brained-qwen3-6-35b-a3b-is-the-fix-b6a687c2094a
9. DeepSeek-V4 Review: Why Million-Token Context Needs Efficient Attention, Not Just Larger Windows, 4月 26, 2026にアクセス、https://artgor.medium.com/deepseek-v4-review-why-million-token-context-needs-efficient-attention-not-just-larger-windows-6dc8e74a00b1
10. DeepSeek V4 Preview Release, 4月 26, 2026にアクセス、https://api-docs.deepseek.com/news/news260424
11. Build with DeepSeek V4 Using NVIDIA Blackwell and GPU-Accelerated Endpoints, 4月 26, 2026にアクセス、https://developer.nvidia.com/blog/build-with-deepseek-v4-using-nvidia-blackwell-and-gpu-accelerated-endpoints/
12. DeepSeek-V4: a million-token context that agents can actually use, 4月 26, 2026にアクセス、https://huggingface.co/blog/deepseekv4
13. DeepSeek is back among the leading open weights models with V4 Pro and V4 Flash, 4月 26, 2026にアクセス、https://artificialanalysis.ai/articles/deepseek-is-back-among-the-leading-open-weights-models-with-v4-pro-and-v4-flash
14. Deepseek v4 models are out and here are benchmarks !(4 versions) - Reddit, 4月 26, 2026にアクセス、https://www.reddit.com/r/LocalLLaMA/comments/1su7o92/deepseek_v4_models_are_out_and_here_are/
15. Compare AI Benchmarks and Tests - LLM Stats, 4月 26, 2026にアクセス、

https://llm-stats.com/benchmarks
16. I benchmarked 7 OpenCode Go models against each other. Flash was the practical default, 4月 26, 2026にアクセス、https://www.reddit.com/r/opencodeCLI/comments/1svzgd4/i_benchmarked_7_opencode_go_models_against_each/
17. Moonshot AI Kimi K2.6 now available on Workers AI · Changelog - Cloudflare Docs, 4月 26, 2026にアクセス、https://developers.cloudflare.com/changelog/post/2026-04-20-kimi-k2-6-workers-ai/
18. Kimi-K2.6 - NGC Catalog - NVIDIA, 4月 26, 2026にアクセス、https://catalog.ngc.nvidia.com/orgs/nim/teams/moonshotai/models/kimi-k2.6
19. moonshotai/Kimi-K2.6 - Hugging Face, 4月 26, 2026にアクセス、https://huggingface.co/moonshotai/Kimi-K2.6
20. Kimi K2.6 - Kimi API Platform, 4月 26, 2026にアクセス、https://platform.kimi.ai/docs/guide/kimi-k2-6-quickstart
21. License - MoonshotAI/Kimi-K2 - GitHub, 4月 26, 2026にアクセス、https://github.com/moonshotai/Kimi-K2/blob/main/LICENSE
22. (Confirmed) Kimi K2's “modified-MIT” license does NOT apply to synthetic data/distilled models : r/LocalLLaMA - Reddit, 4月 26, 2026にアクセス、https://www.reddit.com/r/LocalLLaMA/comments/1m3n89p/confirmed_kimi_k2s_modifiedmit_license_does_not/
23. Qwen3.6 is the large language model series developed by Qwen team, Alibaba Group. - GitHub, 4月 26, 2026にアクセス、https://github.com/QwenLM/Qwen3.6
24. Qwen3.6-27B: Flagship-Level Coding in a 27B Dense Model, 4月 26, 2026にアクセス、https://qwen.ai/blog?id=qwen3.6-27b
25. Qwen3.6 - How to Run Locally | Unsloth Documentation, 4月 26, 2026にアクセス、https://unsloth.ai/docs/models/qwen3.6
26. Gated DeltaNet for Linear Attention - rasbt/LLMs-from-scratch - GitHub, 4月 26, 2026にアクセス、https://github.com/rasbt/LLMs-from-scratch/blob/main/ch04/08_deltanet/README.md
27. Qwen/Qwen3.6-27B - Hugging Face, 4月 26, 2026にアクセス、https://huggingface.co/Qwen/Qwen3.6-27B
28. My key takeaways on Qwen3-Next's four pillar innovations, highlighting its Hybrid Attention design : r/LocalLLaMA - Reddit, 4月 26, 2026にアクセス、https://www.reddit.com/r/LocalLLaMA/comments/1nwzs0k/my_key_takeaways_on_qwen3nexts_four_pillar/
29. Qwen3.6-35B-A3B: Agentic Coding Power, Now Open to All, 4月 26, 2026にアクセス、https://qwen.ai/blog?id=qwen3.6-35b-a3b
30. Xiaomi MiMo-V2.5 Series Large Model Launches Public Beta, 4月 26, 2026にアクセス、https://platform.xiaomimimo.com/docs/news/v2.5-news
31. Xiaomi MiMo-V2.5, 4月 26, 2026にアクセス、https://mimo.xiaomi.com/mimo-v2-5/
32. Xiaomi introduces a new smartphone AI assistant - miclaw - Huawei Central, 4月 26, 2026にアクセス、https://www.huaweicentral.com/xiaomi-introduces-a-new-smartphone-ai-assista

nt-miclaw/
33. MiMo-V2.5-Pro - API, Specs, Playground & Pricing - Puter Developer, 4月 26, 2026にアクセス、https://developer.puter.com/ai/xiaomi/mimo-v2.5-pro/
34. Replies - Community - Xiaomi, 4月 26, 2026にアクセス、https://new.c.mi.com/global/user/39a9109f5033a11225ed9b8be505628a/replies
35. All Specs, Features of Mi AIoT Router AX3600 - Xiaomi, 4月 26, 2026にアクセス、https://www.mi.com/global/product/mi-aiot-router-ax3600/
36. Replies - Xiaomi Community, 4月 26, 2026にアクセス、https://c.mi.com/global/user/089b75de19dbdbcfebf4a013e172c7fb/replies
37. Xiaomi Community | Xiaomi, 4月 26, 2026にアクセス、https://new.c.mi.com/global/user/ad7edb58793514b893a88d884f317875/replies
38. MiMo-V2.5 API — One API 400+ AI Models | AIMLAPI.com, 4月 26, 2026にアクセス、https://aimlapi.com/models/mimo-v2-5
39. Xiaomi Releases MiMo-V2.5-Pro and MiMo-V2.5: Matching Frontier Model Benchmarks at Significantly Lower Token Cost - MarkTechPost, 4月 26, 2026にアクセス、https://www.marktechpost.com/2026/04/22/xiaomi-releases-mimo-v2-5-pro-and-mimo-v2-5-matching-frontier-model-benchmarks-at-significantly-lower-token-cost/
40. I set up Xiaomi's 100% FREE MiMo v2.5 Pro API in VS Code. Is it actually a viable Claude alternative? - Reddit, 4月 26, 2026にアクセス、https://www.reddit.com/r/FastAPI/comments/1surs8z/i_set_up_xiaomis_100_free_mimo_v25_pro_api_in_vs/
41. Best AI for Coding 2026 - Top Coding Models - LLM Stats, 4月 26, 2026にアクセス、https://llm-stats.com/leaderboards/best-ai-for-coding