ChatGPT

生成AIモデル競争の最新動向と今後1〜3ヶ月の予測

競争軸の変化：モデル単体性能から“仕事完遂力”へ

ご提示の記事が示す本質は、**性能競争の軸が「モデル単体の賢さ」から「モデル＋ツール＋エージェントの実行力」へ移りつつある**という点です。Alibaba Cloud（Qwen）やMoonshot AI（Kimi）は、**HLE**（Humanity’s Last Exam）や**BrowseComp**、**SWE-bench Verified**といった「最後までやり切る・探し切る」系ベンチマークで優位性を強調しています。これは、単なる知識量や文章生成力ではなく、「**検索して確かめ、計算し、タスクを完遂する**」一連の能力こそが新たな競争軸になっていることを示唆しています。

特にご要望いただいた**「次の1〜3ヶ月」**および**「マルチモーダル性能、エージェント活用」**に照らすと、今後の競争はさらに以下の点に焦点が移るでしょう：

- **ツール実行の巧拙:** 外部ツール（検索・コード実行・ブラウザ操作など）をいつどのように使い、失敗時にどうリカバリーするか。
- **マルチモーダル入力から成果物への変換:** 画像や動画を入力として受け取り、コードやドキュメントなど**実際に使えるアウトプット**を生成できるか。
- **実運用コスト:** 単位トークン当たりの価格だけでなく、タスク完了までの**試行回数やエラー率**も含めたコスト効率。

Qwen3-Max-Thinking：推論中のツール活用で“考えて調べるAI”に [1] [2]

Alibaba Cloudの「Qwen3-Max-Thinking」（2026年1月23日公開）は、その名の通り**推論特化（Thinkingモード）**のフラッグシップモデルです。最も特徴的なのは、**推論の途中に内蔵ツールを自律的に呼び出せる**点です。具体的には、モデルが回答を生成している最中に必要に応じて**ウェブ検索や情報抽出、コード実行**を自動で挟み込みます [2]。ユーザが指示しなくても、Qwen3-Max-Thinking自身が「ここで検索すべき」「ここで計算すべき」と判断し、検索エンジンやメモリ機能、内蔵のコードインタプリタを呼び出す仕組みです。この“**考えながら調べる**”能力により、事実確認や計算検証を人手を介さず行えるため、**複雑な課題でも途中で手詰まりしにくい**利点があります。

性能面: Alibabaによれば、Qwen3-Max-Thinkingは19項目の代表的ベンチマークでGPT-5.2-ThinkingやClaude-Opus-4.5、Gemini 3 Proと同等の総合性能を示したとされています [3]。特に、**ツール使用あり条件のHLE**では**スコア58.3**を記録し、GPT-5.2-Thinkingの45.5やGemini 3 Proの45.8を大きく上回ったと報告されています [4]。これは**検索ツール等を組み合わせた難問解決**において、Qwenが既存トップモデルを凌駕したことを意味します。また**数学**や**コード推論**でも、高難度試験HMMTで満点近い98.0点を出すなど、従来モデルをリードする結果が示されています [5]。

技術的工夫: Qwen3-Max-Thinkingには**Test-time Scaling**と呼ばれる新手法が導入されており、**推論時に複数ラウンドの自己反省を行って解を洗練**することで、同じ計算量でもより高精度な推論を可能にしています [1]。簡単な質問には速く答え、難しい課題には時間をかけて深く考える「二段階思考」を実現する設計です。このおかげで、**Gemini 3 Proを推論力で上回る**ケースも出てきています [1]。

価格設定: 推論特化モデルでありながら、**利用コストは“高いが手が届く”レベル**と評価されています。実際、Alibaba Cloud国際版でのAPI価格は**入力100万トークンあたり1.20ドル、出力100万トークンあたり6.00ドル**

に設定されています [6] 。これはOpenAIやGoogleの同クラスモデルより安価であり（GPT-5.2では出力$10/百万トークン程度との報告もあり）、**高度な推論を比較的低コストで回せる**ことになります。

短期的展開予測: 今後1〜3ヶ月でQwen側に起こりそうな動きとしては、まず**「OpenAI互換APIで使える高性能エージェントモデル」**という立場を活かし、既存の社内エージェントやRAG（Retrieval-Augmented Generation）システムへの**差し替え導入が進む**と予想されます。Qwen3-Max-ThinkingはOpenAIのAPIと高い互換性を持つため、企業がChatGPT等から乗り換えやすい利点があります。また、ツール自律呼び出し機能の成熟により、**開発者が一からエージェントを構築しなくてもモデル任せでかなり動く**状況が生まれつつあります。短期の競争ポイントは、こうした**ツール呼び出しの安定性**です。検索やコード実行でエラーが起きた際、モデルがいかに**上手にリトライ・別経路探索できるか**が実務上の評価を左右するでしょう。

Kimi K2.5：ネイティブ・マルチモーダルとエージェント群による“何でも屋”モデル [7] [8]

Moonshot AIの「Kimi K2.5」（2026年1月27日公開・オープンソース）は、**テキストに加えて画像や動画をネイティブに扱えるマルチモーダル**大規模モデルです [9] 。15兆個もの視覚・言語混合トークンで事前学習されており、**UIデザインのスクリーンショットやワークフローの動画を入力に、対応するコードやアプリを生成**することすら可能と謳われています。MoonshotはK2.5を「現時点で最強のオープンソースモデル」と位置づけており、その性能を示すため**複数の指標（HLE、BrowseComp、MMMU-Pro、SWE-bench Verified など）**のスコアを公開しています。特に**エージェント系タスク**（自律的な問題解決）では、GPT-5.2やGemini 3 Proを上回る成績を収めたとアピールしています [10] 。例えば**コードベンチマーク（SWE-Bench Verified）**ではKimi K2.5がGemini 3 Proを凌ぎ、**動画理解**ではGPT-5.2やClaude-Opus-4.5をも上回ると報じられています [10] 。

最大の特徴: Kimi K2.5が業界に与えたインパクトは、**「エージェント・スウォーム（Agent Swarm）」**という新しいパラダイムです [11] 。これは単一のエージェントではなく、**モデル内部で多数のサブエージェントを並列稼働させる**仕組みです。Moonshotによれば、K2.5は**最大100体のサブエージェント**を自律的に編成し、**最大1500回ものツール呼び出しを並行実行**できます [8] 。これにより、従来の1体エージェントによる処理と比べて**最大4.5倍の速度向上**を達成し得るとされています [8] 。注目すべきは、これらのサブエージェントや作業ワークフローを**事前に人間が定義せずともモデル自ら自動生成・オーケストレーションする**点です [8] 。従来、多エージェントの協調は外部フレームワークで試行錯誤する必要がありましたが、Kimiはそれ自体をモデル学習に取り込む方向で差別化を図っています。

周辺ツール: また、K2.5の能力を活かすため**「Kimi Code」**と呼ばれるオープンソースのコーディング支援ツールも公開されました [12] 。AnthropicのClaude CodeやGoogleのGemini CLIに対抗する位置付けで、ターミナルやVSCode、Cursor、Zedなど開発環境と統合して使えます [12] 。画像や動画を入力し、それを元にしたコード生成やUI構築まで可能で、**エンジニアの実務フローに深く入り込む**ことを狙ったツール群です。

短期的展開予測: 今後1〜3ヶ月では、Kimi K2.5はまず**「Agent Swarm」の実利用拡大**を目指すでしょう。現在ベータとされている並列エージェント機能が、より多くのユーザプランやオープンソースコミュニティで試せるようになる可能性があります。また**Kimi CodeのIDE連携**（プラグインや拡張機能の提供）が進み、実際の開発現場で使われる機会を増やすと予想されます。競争ポイントとしては、**多数のサブエージェント間で生じる矛盾をどう解消するか、1500回にも及ぶツール呼び出しによるコスト増・失敗率増をどう抑えるか**、そして**ログやコンテキストの管理**（大量の中間結果をどう要約・共有するか）といった点が挙げられます。短期的には、これら「実装力」が高いチーム・モデルほど評価を上げるでしょう。

ベンチマーク動向：“ツール完遂型”評価へのシフト

モデル評価の文脈でも、**何を測るか**が変化しつつあります。従来は学術試験や知識問答（MMLUや数学コンペなど）で正答率を競う場面が多かったですが、QwenやKimiが強調するHLE・BrowseComp・SWE-benchといった指標は**ツール使用やエージェント行動を前提**としています。いくつか主要ベンチマークの特徴を整理します：

- **HLE (Humanity’s Last Exam):** 各分野の専門家が作成した難問テスト。選択肢問題だけでなく**図表の読解や検索を要する問い**も含む総合評価です。モデルの純粋知識だけでなく、**外部情報を探してでも正解に辿り着けるか**を測る面があります。
- **BrowseComp:** Webブラウジングを伴う探索課題です。インターネット上の隠れた情報を見つけ出す力（適切なクエリ構築、情報精査、途中での方向転換など）を評価します。モデルが**どれだけ粘り強く目的の情報に到達できるか**という、実務的なリサーチ能力の指標です。
- **MMMU / MMMU-Pro:** マルチモーダル（テキスト＋画像や動画）で多分野の大学レベル問題を解くテスト。MMMU-Proでは**定型解答やショートカットを排除**し、モデルが本当に理解しているかを厳密に判定します。
- **SWE-bench Verified: 実コードの課題を解決し、単体テストをすべて通過させる**競技的ベンチマークです。与えられたバグ修正や機能実装タスクに対し、モデルがコードを書き、テストに合格するところまで評価します（Verifiedは人手確認ずみの正解があるサブセット）。

記事にもあったように、今後1〜3ヶ月の競争では**HLEのような「静的な難問での正答率」よりも、BrowseCompやSWE-bench　Verifiedのような「ツールを駆使してタスクを最後まで完遂する率」**が重視される見込みです [4] 。つまり、**一問一答の賢さ**以上に、**長時間・複数ステップにまたがるタスクを切り抜ける粘り強さ**や**完了精度**が評価指標の中心に据えられるでしょう。

今後1〜3ヶ月の競争予測：マルチモーダル×エージェントの行方

以上を踏まえ、直近数ヶ月（2026年2月〜4月）の生成AI競争について、特に**マルチモーダル対応**と**エージェント活用**の観点から予測される動向を整理します。

予測A：競争軸は「モデル性能」から「エージェント実行環境」へ
モデルそのものの賢さよりも、**そのモデルをどう“働かせる”か**が勝敗を分ける局面になりそうです。具体的には、各社が提供する**エージェント用プラットフォーム**や**統合開発環境（IDE）**が重要になります。例えばGoogleはエージェント前提の開発環境「Antigravity」を発表しており、複数のエージェントが連携してコードを書き、進捗をArtifacts（成果物ログ）として見せる仕組みを推しています [13] [14] 。このように、単に「どのモデルが一番頭が良いか」ではなく**「どの環境ならAIが人手を減らして仕事を片付けてくれるか」**が企業ユーザの関心になります。したがって短期的には、モデル単体の微差よりも**ツール連携の種類の豊富さ、エラー発生時のリカバリ自動化、実行ログや差分の可視化（監査しやすさ）**といった要素で競争が展開されるでしょう。

予測B：マルチモーダルは「派手な生成」より「画面・動画から作業完遂」へ
画像生成そのものの品質競争は既に一定水準に達しつつあり、次の焦点は**「視覚情報を使ってタスクを完了する」**ことに移ると考えられます。Kimi　K2.5が打ち出したように、**UIのスクリーンショットや操作動画→コード生成・UI再現**という流れが一例です [15] 。Gemini　3も「テキスト・画像・音声・動画・ドキュメント」を横断的に理解する能力を掲げており、例えば**画面キャプチャから必要な変更点を読み取ってコード修正し、結果をプルリクエストとして提出**するといった高度な自動化が目標に据えられています。OpenAI（GPT-5.2）もリリースノートで「日常業務タスクへの対応強化」（スプレッドシートの自動処理やスライド作成補助など）に言及しています。したがって、2〜4月にかけては**「画像/動画など視覚的資料を工程に組み込んだ実務タスク」でどのモデルが最もスムーズに完遂できるか**が顕著な差別化ポイントになるでしょう。

予測C：安全規制が“攻めの機能”にブレーキをかける可能性（特にGrok）
Elon Musk率いるxAIの**Grok**は、画像生成・編集機能に関連して早くも**規制当局の強い視線**を浴びています。非同意の性的ディープフェイク生成問題で、EUが正式調査を開始し（DSA違反の可能性）[16] [17]、イギリスでもオンライン安全法に基づく調査・措置が進行中です。英国政府は「性的ディープフェイクは悪質な虐待の武器であり、プラットフォーム提供自体を違法とする」方針を打ち出しました[18]。短期的にGrokはこの問題対応に追われ、**画像生成機能の制限強化やガードレール改善**が最優先課題となるでしょう。結果として、Grokが本来アピールしたかった**攻めのマルチモーダル機能**（例えば高度な画像編集や合成）は一時的に影を潜める可能性があります。他方、OpenAIやGoogle、Anthropicといった他社は「**企業が安心して使えるマルチモーダル**」を追求することで、中長期的に信頼性の面でリードを広げる戦略を取ると考えられます。

主要モデルの近々のアップデート情報（2026年2〜4月）

次に、主要な生成AIモデル（ChatGPT、Gemini、Claude、Grok）の今後1〜3ヶ月で予想されるバージョンアップや新機能動向をまとめます。各社とも年末〜年始に大型アップデートを行った直後であり、それを踏まえた展開となります。

- **ChatGPT（OpenAI）：** 2025年12月に**GPT-5.2**ファミリー（Instant / Thinking / Pro）が公開され、ChatGPTの既定モデルも順次5.2系に切り替わりました[19]。旧GPT-5系モデルは“レガシー”扱いで、5.2ローンチ後3ヶ月程度（~2026年3月頃）まで提供継続という情報もあります。短期的には、新ナンバリング（GPT-5.3等）よりも**ツール実行の安定化**や**エージェント機能（Code Interpreterやブラウザプラグイン）の改善**が中心になるでしょう。OpenAIは「自社ツール（ブラウザ検索など）を組み合わせたHLE成績でGemini 3 Proを上回った」とも伝えられており[20]、引き続き**内蔵ツール＋大規模モデル**の完成度を高めるアップデートが予想されます。例えば、長時間の対話でログが肥大化した際の**要約引き継ぎ**や、エージェントの**タスク再開**機能の強化など、実務利用での痒い所に手を届かせる改良が期待できます。

- **Gemini（Google DeepMind）：** 2025年末に**Gemini 3**が登場し、最上位のGemini 3 **Pro**が一般提供されました。画像生成専用の**Nano Banana Pro**（Gemini 3由来の高性能画像モデル）や、超高機能推論モードの**Deep Think**（Ultraプラン向け）も発表されています[21] [22]。Gemini 3 Pro自体はLMArenaベンチマークで1501 Eloを記録し、従来トップのGrok 4.1 Thinkingを上回るなど高評価です[23]。**Deep Think**モードについては「近日中にUltraユーザー向け提供」とティーザーされており[22]、今後数ヶ月で段階的に公開されるでしょう。またGoogle検索（AIモード）へのGemini統合も進んでおり、**検索クエリが難しいと判断した場合に自動でGeminiに振り分ける**機能が追加予定です[24]。こうした大規模分散インフラを持つGoogleの強みは、Geminiを**エンタープライズや消費者向けサービスのデフォルトAIエンジン**に押し上げる可能性があります[14]。2〜4月にかけては、**Google WorkspaceやAndroidへのGemini導入**、**サードパーティ向けVertex AIでのGemini提供拡大**なども想定され、幅広いユースケースでGeminiモデルを利用できる環境が整備されていくでしょう。

- **Claude（Anthropic）：** Anthropicは2025年末に**Claude 4.5**（Opus〈Max相当〉やSonnet〈Pro相当〉）をリリースし、その派生として**Claude Desktop**アプリで動作する汎用エージェント機能「**Claude Cowork**」を2026年1月に研究プレビュー公開しました[25]。当初はMaxプラン限定でしたが、1月16日にはProプラン($20/月)にも拡大されています[26]。Coworkは「Claude Codeの非開発タスク版」という位置づけで、ユーザがフォルダを指定するとその中のファイルをスキャンし、必要に応じて**ローカルでコードやコマンドを実行しながら**タスクを遂行します[27] [28]。さらにAnthropicは**コネクタ（Connectors）**と**エージェント・スキル（Agent Skills）**を2026年1月に全ユーザー（Pro含む）へ一般提供しました[29]。コネクタは外部のSaaSやデータベースにClaudeを接続する機能、スキルは特定業務（例：医療のFHIRデータ処理など）向けのツールセットです。短期的なアップデートは、Coworkの改善サイクル（Anthropicは「迅速な改善」を公言[25]）に沿って**対応アプリの増加**（例: Windows版リリースやSlackなど他プラットフォーム統合）、**セキュリティ対策強化**（誤操作防止やプロンプトインジェクション耐性）などが中心になるでしょう。また、2026年初頭には**Excel用**

Claudeや**Chrome用Claude拡張**も登場しており、Anthropicは「日常業務に溶け込むAIエージェント」としてClaudeを位置づけています。よって今後数ヶ月も、**ビジネスソフトとの直接連携**や**業務特化スキルの拡充**にフォーカスしたアップデートが予想されます。

- **Grok（xAI）:** xAIの**Grok 4**は2025年秋、改良版の**Grok 4.1**は同年11月に公開されました（パラメータは非公開ながら推定数千億〜1兆規模）。Grok 4.1はリリース直後にLMSYSリーダーボードでトップクラスの評価を得ていましたが [30]、12月以降Gemini 3やClaude 4.5の登場で僅差ながら追い抜かれています [23]。2026年1月現在、xAIは企業向けに**ビジネス/エンタープライズ版**を提供し、**社内アプリ接続**や**カスタムエージェント**機能を売り込み中です。しかし前述のとおり、**Grokの画像生成機能に対する各国規制当局の調査**が相次いでおり、短期的には**機能制限や安全対策強化**が避けられません [16] [18]。欧州委員会はGrokのリスク評価義務違反の疑いで調査を開始し、違反と認定されれば世界収益の6%までの罰金もあり得る状況です [16] [17]。イギリスもオンライン安全法の初適用ケースとしてX（旧Twitter）社に圧力を強めており [31]、「サービス提供停止（プラットフォームの国内遮断）の可能性も排除しない」としています [32]。こうした背景から、今後1〜3ヶ月でxAIが積極的な新機能（例: Grok 4.2や新たな生成能力）を打ち出す可能性は低く、むしろ**既存機能の絞り込み**や**不適切用途へのガードレール実装**がアップデートの中心となるでしょう。コミュニティで噂される「Grok 4.2」の存在も確証はなく、当面は公式の安全対応発表に注目すべき段階です。

短期決戦のカギ：「壊れずにやり切る」AIはどれか？

最後に、今後1〜3ヶ月の競争で**実務ユーザが評価するポイント**を整理します。各社モデルの差異は細かなベンチマークスコアにも表れますが、実際に業務に投入する際には以下のような観点が勝敗を左右します：

- **タスク完遂率:** 例えば10分〜30分かかる複合作業（複数ツールを順番に使うようなタスク）を、一度の指示でどこまで完了できるか。途中で手動介入が必要になる割合が低いモデルが好まれます。
- **リトライやフォローの回数:** 一度で完璧にいかなくても、モデル自身がミスに気付き再試行したり、追加情報を取得して軌道修正できるか。人間が追加プロンプトを与えて調整する手間が少ないほど高評価です。
- **ツール呼び出し効率:** エージェントが**何回外部ツールを呼ぶか**はコストとリスクに直結します。呼び出しが多すぎればAPI費用が膨らみますし、エラーの発生確率も上がります（特にKimiのように最大1500回となると尚更です）。必要十分な回数で目的を達成できるバランス感覚が問われます。
- **監査性（証跡の残し方）:** モデルが**どの情報を参照し、どんなコマンドを実行し、何を変更したか**のログが明確に残るかも重要です。業務利用では結果の正確さだけでなく、プロセスをあとからレビューできることが求められます。ログや差分、引用元URLなどの「証跡」を自動生成・提示できる機能は大きな強みになります。

こうした観点で見たとき、**「途中で壊れず最後まで走り切る」**モデルが短期的な評価を勝ち取るでしょう。Qwen3-Max-ThinkingやKimi K2.5が強調する“**モデル＋ツール＋エージェント**”のアプローチはまさにその方向性であり、各社の開発競争もここに集中しています。2〜4月の動向として、**ユーザ企業は自社の具体的タスクをモデルにやらせてみて、その完遂コストで採用を判断する**ケースが増えると考えられます。言い換えれば、「一問一答の正確さ」より「一連の仕事を任せたときの安心感・コストパフォーマンス」が評価基準となり、そこを制したモデル/プラットフォームが短期的な競争をリードするでしょう。

以上、**中国勢の最新モデル（Qwen3-Max-ThinkingとKimi K2.5）**の動向を軸に、主要モデルの近未来アップデート情報とマルチモーダル×エージェント競争の予測を深掘りしました。各種情報は現時点で公開されたソースに基づいておりますが、生成AI業界は日進月歩であり、今後も新たな発表に注視が必要です。 [4] [10]

---

1 3 5 Qwen3-Max-Thinking Outperforms Claude Opus 4.5 and Gpt-5.2: The Panic Is Real | by Agent Native | Jan, 2026 | Medium
https://agentnativedev.medium.com/qwen3-max-thinking-outperforms-claude-opus-4-5-and-gpt-5-2-the-panic-is-real-d4c2557879d4

2 6 I replaced ChatGPT with Alibaba's new reasoning model for a day — here's what Qwen3-Max-Thinking does better | Tom's Guide
https://www.tomsguide.com/ai/i-replaced-chatgpt-with-alibabas-new-reasoning-model-for-a-day-heres-what-qwen3-max-thinking-does-better

4 阿里深夜发布：号称最强千问推理模型，比肩GPT-5.2
https://m.mp.oeeee.com/a/BAAFRD0000202601271515173.html

7 9 10 12 15 China's Moonshot releases a new open source model Kimi K2.5 and a coding agent | TechCrunch
https://techcrunch.com/2026/01/27/chinas-moonshot-releases-a-new-open-source-model-kimi-k2-5-and-a-coding-agent/

8 11 Kimi K2.5: Visual Agentic Intelligence
https://simonwillison.net/2026/Jan/27/kimi-k25/

13 14 22 23 24 Google's new Gemini 3 Pro is a great model - Creative Strategies
https://creativestrategies.com/research/gemini-3-pro-is-a-great-model/

16 17 X faces EU investigation over Grok's sexualized deepfakes | The Verge
https://www.theverge.com/news/868239/x-grok-sexualized-deepfakes-eu-investigation

18 31 32 UK investigates Musk's X over Grok deepfake concerns | Reuters
https://www.reuters.com/business/media-telecom/uk-regulator-launches-investigation-into-x-over-grok-sexualised-imagery-2026-01-12/

19 GPT-5.2 - Wikipedia
https://en.wikipedia.org/wiki/GPT-5.2

20 国産技術で宇宙の暗闇を照らす…K-DRIFT望遠鏡、超微光の初観測に ...
https://finance.biggo.jp/news/nwDT_psBUUDt0E6ps9S0

21 Introducing Nano Banana Pro - Google Blog
https://blog.google/innovation-and-ai/products/nano-banana-pro/

25 26 27 28 First impressions of Claude Cowork, Anthropic's general agent
https://simonwillison.net/2026/Jan/12/claude-cowork/

29 Advancing Claude in healthcare and the life sciences \ Anthropic
https://www.anthropic.com/news/healthcare-life-sciences

30 Grok Changelog
https://grok.com/changelog