AI MODEL TECHNICAL & MARKET REPORT | 2026.11.16

Alibaba「Qwen4-Max」徹底調査レポート

2.4 兆パラメータ新世代 MoE モデルの技術仕様・客観的ベンチマーク評価・国内外コミュニティの評判

Gemini 4 Carbon

対象モデル	Qwen4-Max (GA 正式版)	**開発・提供元**	Alibaba Cloud / Qwen Team
正式リリース日	2026 年 11 月 16 日	**調査作成日**	2026 年 11 月 16 日（参考文献 12 件収録）

エグゼクティブ・サマリー（調査要旨）

2026 年 11 月 16 日、中国 Alibaba Cloud（Qwen 開発チーム）は次世代フラッグシップ AI モデル「**Qwen4-Max**」を正式リリースした[1, 3]。総パラメータ数 **2.4 兆（活性化 710 億）**の新設計 Sparse MoE アーキテクチャと **100 万トークン**のコンテキストウィンドウを備え、前世代（Qwen3.8-Max）比でタスクあたりの実効コストを **62%削減**している[1, 4]。長期ソフトウェア開発（DeepSWE v1.1）や実務エージェント評価（Agents' Last Exam v1.1）では Anthropic の Claude Opus 5.5 を上回るスコアを達成した一方[1, 5]、ターミナル操作（Terminal-Bench 4.0）や総合知能指数（Artificial Analysis: 43 点・38 位）では米中の一部最上位モデルに一歩譲る結果となった[6, 7, 10]。開発者コミュニティからは「**圧倒的なトークン効率・長文キャッシュ単価の安さ・視覚能力の改善**」が高く評価されており、後日公開が予告された小型オープンウェイト版にも大きな注目が集まっている[1, 10, 11, 12]。

1. リリースの経緯と位置づけ

2026 年 11 月 16 日、Alibaba（アリババ）傘下の Qwen 開発チームは公式ブログにおいて、*“Qwen4-Max: Your New Teammate Has Arrived”* と題して第 4 世代 Qwen シリーズの最上位モデル「**Qwen4-Max**」の一般提供（GA）開始を発表した[1, 5]。同モデルは、中国の大規模商戦「独身の日（ダブル 11）」にあたる 2026 年 11 月 11 日からチャットサービスおよび定額制プラン上で先行公開されていたプレビュー版（`Qwen4-Max-Preview`）を経て[8, 9]、アーキテクチャの技術詳細、各種ベンチマーク結果、および従量課金 API の正式料金体系とともに公開されたものである[1, 2, 3]。

Qwen チームは同モデルを単なる対話型チャットボットではなく、**「仕事・創造・科学的発見のための新しい AI チームメイト」**と定義している[1]。26 の Tier-1 ドメインおよび 721 の専門分野を網羅する実務ベンチマーク「Work360」の開発・適用をはじめ、3D 制作・ゲーム開発・精密製造・金融・法務・臨床医療といった産業ワークフローへの直接統合を強く打ち出している点が最大の特徴である[1, 3]。また、Qwen4-Max 自体はクローズドな API・クラウド経由での提供となるが、**同一の Qwen4 アーキテクチャおよび学習技術を用いた小型モデルのオープンウェイト（オープンソース）公開**が公式に予告されている[1, 5, 12]。

2. 技術仕様と新世代アーキテクチャの革新

■ 2.1 2.4 兆パラメータ「Sparse MoE」と 3 つの新機構

Qwen4-Max は刷新されたスパース MoE（Mixture-of-Experts）構造を採用しており、バックボーンの総パラメータ数は **2.4 兆（2.4T）**、1 トークン推論あたりの活性化パラメータ数は **710 億（71B）**に設計されている[1, 4]。前世代のベースモデル（`Qwen3.8-Max-Base`）と比較して活性化パラメータを 25%削減しながら、知識・コード・品質保証（QA）・STEM の全評価領域において BPB（Bits Per Byte：低いほど優秀）を約 5〜13%改善した[1, 3]。事前学習では以下の 3 つのアーキテクチャ革新が導入され、2.4T 規模の学習をロススパイク（損失の急騰による不安定化）なしで完遂させている[1, 4]。

- **GDN + Qwen Sparse Attention (QSA)**：3 層の Gated DeltaNet（GDN）層と 1 層の QSA 層を組み合わせたハイブリッド機構。GDN が文脈履歴を固定サイズの状態へ圧縮し、QSA の軽量インデクサーがマイクロブロック単位で関連性の高いコンテキストのみを選択的に参照することで、長文処理時のアテンション計算・インデックス負荷を大幅に削減する[1, 3, 4]。
- **Gated Residual (GR)**：動的な読み書きゲートを備えた 4 つの並列残差（Residual）分岐を導入。層間の情報伝達力と表現力を高めつつ、超大規模パラメータ学習時の勾配安定性を確保している[1, 3, 4]。
- **N-gram Embedding（追加 3,280 億パラメータ）**：モデル第 2 層において局所的な連続トークンパターンの表現を直接引き当てる埋め込み機構を採用。推論時の演算コスト（FLOPs）をほとんど増加させることなく、モデル全体の事実知識容量を拡張している[1, 3, 4]。

■ 2.2 マルチモーダル対応・100 万コンテキスト・推論制御

- **ネイティブ視覚理解（Visual Intelligence）**：テキストだけでなく画像・動画のネイティブ入力に対応し、テキストを出力するマルチモーダル基盤を備える[1, 3, 5]。UI レイアウトの視覚的検証や 3D 空間・映像シーンの理解精度が前世代から大幅に引き上げられた[1, 3]。

- **入出力トークン上限**：最大 **100 万（1M）トークン**のコンテキストウィンドウをサポートし、出力上限は思考プロセスと最終回答の合算で最大 **262,144 トークン**に対応する[1, 2, 4]。
- **推論（Reasoning）モードの柔軟化**：前世代の Qwen3.8-Max では思考モードを完全に無効化できなかったが、Qwen4-Max では`none`（推論なし）から`low / high / xhigh`までタスクの難易度とレイテンシ要件に応じた段階的制御が可能となった[1, 3]。

■ 2.3 トークン効率の劇的改善と API 料金体系

冗長な思考トークンを抑制する学習最適化により、前世代（Qwen3.8-Max）比で「Work360」評価時のトークン消費量を約 24%（xhigh 設定）〜35%（high 設定）削減、「Terminal-Bench 4.0」では約 48%〜49%削減した[1, 3]。さらに API 単価の引き下げと 95%割引のコンテキストキャッシュ料金が組み合わさることで、**Work360 における 1 タスクあたりの平均実行コストは 4.43 ドルから 1.67 ドルへと 62%低下**している[1, 3]。

表 1：Qwen4-Max 主要技術仕様および API 料金一覧[1, 2, 3, 4, 5, 6]

項目	仕様・数値	備考・前世代（Qwen3.8-Max）比の特徴
モデル構造	Sparse MoE（新 Qwen4 世代）	GDN+QSA、Gated Residual（4 分岐）、N-gram Embedding 搭載[1, 4]
パラメータ規模	総数 2.4 兆（2.4T） / 活性 710 億（71B）	N-gram Embedding 用に別途 3,280 億パラメータ、活性化数は前世代比 25%減[1, 4]
対応モダリティ	テキスト・画像・動画入力 → テキスト出力	ネイティブ画像認識・映像フィードバックループに対応[1, 5]
コンテキスト長	入力 最大 1,000,000（1M）トークン	出力（思考＋回答）最大 262,144 トークン[2, 4]
推論制御モード	none / low / medium / high / xhigh	非推論（none）モードが新設され低遅延応答に対応[1, 3]
API 入力料金	$1.50 / 100 万トークン（約 239 円）	100 万トークン上限まで長文割増なしのフラット料金[2, 5, 6]
API 出力料金	$4.50 / 100 万トークン（約 718 円）	業界中央値（$10.00/1M）を大きく下回る価格設定[5, 6]
キャッシュ入力料金	$0.075 / 100 万トークン（約 12 円）	通常入力比 95%割引、反復エージェント処理に最適化[1, 5, 6]

項目	仕様・数値	備考・前世代（Qwen3.8-Max）比の特徴
提供チャネル	QwenCloud API / Token Plan / Chat 等	QwenWork、Qoder、Qwen Studio、chat.qwen.ai で利用可能[1, 3, 9]

3. 実務応用・24/7 エージェント機能・セキュリティ強化

公式発表および技術報道では、単体ベンチマーク以上に「複雑な実務ソフトウェア環境での自律遂行力」が強調されている[1, 3, 5]。

- **長時間稼働エージェント（24/7 Agents）と「Live Delivery」**：自律的な計画立案と継続的な反復修正によって長時間タスクを遂行し、中間成果物を段階的に納品しながらユーザーがいつでも介入・軌道修正できる「Live Delivery」機能を搭載した[1, 3]。あわせて専用プラグイン`Qwen-Live-Delivery-Plugin`もオープンソース公開されている[1]。
- **3D 制作・ゲーム開発・映像生成ワークフロー**：Blender を直接操作して 3D プリント可能な精密 3D モデルやシーンを構築できるほか、Unity 6、Unreal Engine、Godot Engine と連携し、実際にゲームを実行してカメラ追従・当たり判定・ライティングを視覚フィードバックで反復改善しながら 3D アクションやモバイルレースゲームを完成させる事例が公開された[1, 3, 5]。動画分野でも WonderClip との連携により絵コンテ作成から複数シーンの映画編集までを一貫処理できる[3]。
- **精密製造（CAM）・専門職（金融・法務・医療）・科学探索**：自動車部品加工の検証において過切削や削り残しを事前検知し、ワークピース 1 件あたりの CAM プログラミング手作業時間を**約 475 分から 24 分へと約 95%短縮**した[1, 3]。また、90 種の MCP（Model Context Protocol）ツールを駆使した金融投資調査、IPO 前の法務デューデリジェンス、臨床医療の意思決定支援、さらにはデータセンター冷却向け新材料の自律設計・合成探索にも適用されている[1, 3]。
- **コードセキュリティと安全対策の向上**：ソフトウェア開発学習タスクにコードセキュリティ検証を組み込んだ結果、機能性と安全性を両立する関数レベルの合格率が前世代の 22.0%から **31.8%**へ上昇し、リポジトリ単位で安全なコードを生成する割合も 8.0%から **14.7%**へとほぼ倍増した[3]。CBRN（化学・生物・放射性物質・核）やサイバーセキュリティ領域でも専用ポストトレーニングと内部専門家による CoT（思考連鎖）監査が実施されている[1, 3]。

4. ベンチマークおよび第三者機関による客観的評価

公式発表データおよび独立系評価機関「Artificial Analysis」等の外部測定結果を総合すると、Qwen4-Max の性能プロファイルは「**長期ソフトウェア開発・Web 構築・専門知識タスクでは世界トップ層に食い込**

むー方、ターミナル操作や総合知能リーダーボードでは米中の最上位モデル群を追う立場にある」と整理できる[1, 5, 6, 7, 10]。

表 2：Qwen4-Max 主要ベンチマーク・外部リーダーボード評価一覧[1, 3, 5, 6, 7, 10]

評価ベンチマーク / 指標	Qwen4-Max	主な競合比較・順位・評価ポイント
DeepSWE v1.1（長期 SW 開発）	**75.7**	Claude Opus 5.5 や GPT-6 Astra を上回り比較対象中トップ[1, 5]
Agents' Last Exam v1.1（実務 Agent）	**46.7**	Claude Opus 5.5（43.4）を抑えて首位を記録[1]
Work360（26 領域・721 専門実務）	**52.69**	前世代 Qwen3.8-Max（47.74）から+4.95pt 向上、コスト 62%減[1, 3]
Code Arena | WebDev（Web 開発）	**世界 6 位**	Anthropic「Claude Fable 5.1」を上回る高順位を獲得[5]
Agent Arena | Capabilities	**世界 14 位**	「Claude Sonnet 5.5」および「GPT-6 Sol」より上位にランクイン[5]
Terminal-Bench 4.0（端末操作）	**48.5**	Claude Opus 5.5（59.6）や中国競合 Kimi K3.1（67.2）に劣後[1, 10]
JobBench（職務遂行タスク）	**68.5**	Claude Opus 5.5（75.4）を下回るスコア[1]
Artificial Analysis 知能指数 v5.0.1	**43（38 位/81）**	中央値（41）や Claude Haiku 5.5 は上回るが、Claude Fable 5.2 や Gemini 4 Carbon、Kimi K3.1 Pro には差をつけられる[5, 6, 7]
AA タスクあたり平均コスト	**$3.27（44 位/80）**	良好な経済性。出力/入力比率は 1.284%（冗長性ランク 65 位/80 と簡潔）[6]

また、法務評価（HLAB-Vals-1.1）や生命科学（BioMysteryBench）などの専門知識領域で高い到達度を示した一方、大規模な自動コードシミュレーション計算量を要する`Terminal-Bench 4.0`や`ProgramBench`で伸び悩んだ点について、海外の AI 研究者・開発者からは「米国の輸出規制下における中国ラボの計算資源（コンピュート）制約が、事前学習よりも強化学習シミュレーション工程に影響しているのではないか」という構造的な分析も提示されている[7, 10]。

5. 国内外の開発者コミュニティ・メディアの評判

先行プレビュー（11 月 11 日）から正式リリース（11 月 16 日）にかけて、Reddit（`r/singularity`、`r/LocalLLaMA`、`r/LocalLLM`）や国内外のテックメディアで活発な実機検証と議論が交わされている[5, 8, 10, 11, 12]。

■ 5.1 ポジティブな評価（歓迎されている点）

- **トークン効率とキャッシュ価格による圧倒的コストパフォーマンス**：無駄な思考ループが減って出力が引き締まり（Artificial Analysis でも出力の簡潔さが高評価）[6]、さらに「100 万トークンまで単価据え置き」「キャッシュ入力$0.075/1M」という価格設定が、巨大コードベースや大量文書を繰り返し参照するエージェント運用で極めて実用的だと歓迎されている[1, 5, 10]。
- **フロントエンド・3D 生成力と視覚理解の大幅な改善**：前世代（Qwen3.8 系）で弱点と指摘されていた画像認識や UI 崩れの検知能力が大きく向上した[1, 3]。GIGAZINE によるプレビュー版の実機検証でも、豪華なエフェクトを備えた 3DCG ブロック崩しゲームを単一 HTML として一発で動作可能な品質で生成できることが確認されている（思考プロセスは主に中国語で展開）[8]。
- **中価格帯コーディング・実務モデルとしてのバランスの良さ**：Artificial Analysis、Vals、LiveBench の 6 種のエージェント・コード指標を統合したコミュニティ集計では、`GPT-6.1 Sol Medium`（72.9%）を上回る 73.6%を記録し、コストあたりの実務性能で高く評価する声が上がっている[10]。

■ 5.2 ネガティブ・慎重な評価（指摘されている課題）

- **絶対的な知能上限では最上位フロンティアに届かないという見方**：コミュニティでは「Qwen3.8 登場時ほどのゲームチェンジャー感はない」「体感知能は Claude Opus 5.5 の Medium 設定前後であり、Claude Fable 5.2、Gemini 4 Carbon、GPT-6.1 Terra Max、あるいは中国勢の Kimi K3.1 Pro を凌駕する水準ではない」との冷静な評価が目立つ[5, 7, 10, 11]。手離れの良いシニア SWE 代替というよりは、適切な人間のガイダンス（Live Delivery 等）と組み合わせて回す高効率モデルと受け止められている[10]。
- **プレビュー期の仕様上の癖とサービング安定性への注視**：2.4T という巨大モデルゆえのスループットや、プレビュー版で見られた「思考トークンと回答トークンの合算上限設定を小さくすると回答前に切断される仕様」「サーバー側注入ルールに起因する過剰なプロンプトインジェクション自己防衛ループ」などが本番 API でどこまで解消・安定化しているか注視する声がある[8, 11]。
- **ローカル LLM 勢の最大関心は「今後の小型オープンウェイト版」へ**：`r/LocalLLaMA`等のオープンソースコミュニティでは、クローズド API である Qwen4-Max そのものよりも、**「今回実証された GDN+QSA や N-gram Embedding の新設計と改善された視覚能力を引き継ぐ小型オープンウェイト**

モデル（27B クラス等と目される後続モデル）がいつ公開されるか」に最大の期待が寄せられている[1, 10, 11, 12]。

6. 総括と今後の展望

2026 年 11 月 16 日に正式リリースされた **Qwen4-Max** は、あらゆるベンチマークで世界首位を独占する「汎用知能の単独王者」を目指すモデルというよりも、「**新設計の 2.4T Sparse MoE アーキテクチャによってトークン消費量と API コストを劇的に圧縮し、100 万トークンの長大文脈とネイティブ視覚認識を武器に Web 開発・3D 制作・産業ワークフロー・長時間エージェント業務を高い経済性で遂行する実務特化型フラッグシップ**」として明確な独自ポジションを確立したといえる[1, 3, 6]。

今後は、（1）本番 API 環境における長文・高負荷時の推論速度と安定性、（2）90 種の MCP ツールや`Qwen-Live-Delivery-Plugin`を活用したエンタープライズ導入事例の広がり、そして（3）公式に予告された **Qwen4 世代の小型オープンウェイトモデル群**のリリース時期とローカル環境での性能到達点が、AI 業界および開発者コミュニティにおける最大の注目焦点となる[1, 3, 5, 12]。

参考文献 （References）

本レポートの作成にあたり参照した一次情報（公式発表・API ドキュメント）、第三者ベンチマーク機関、報道メディア、および開発者コミュニティの出典一覧は以下の通りである（すべて 2026 年 11 月 16 日時点で確認）。

[1] **Qwen Team (Alibaba Cloud),** “Qwen4-Max: Your New Teammate Has Arrived”, Qwen Official Blog, 2026 年 11 月 16 日. https://qwen.ai/blog?id=qwen4-max

[2] **QwenCloud (Alibaba Cloud),** “Qwen4-Max - Model Overview, Context Window & API Pricing”, QwenCloud Documentation, 2026 年 11 月 16 日. https://www.qwencloud.com/models/qwen4-max

[3] **AI Watch（Impress Watch）,** 「2.4 兆パラメータの新 AI『Qwen4-Max』を提供開始」, 株式会社インプレス, 2026 年 11 月 16 日. https://ai.watch.impress.co.jp/docs/news/2156788.html

[4] **MarkTechPost,** “Alibaba Releases Qwen4-Max: 2.4T MoE, 71B Active, 1M Context”, MarkTechPost, 2026 年 11 月 16 日. https://www.marktechpost.com/2026/11/16/alibaba-qwen4-max-release/

[5] **GIGAZINE,** 「『Qwen4-Max』が登場、一部テストで Claude Opus 5.5 に勝利し小型オープンモデルも後日公開予定」, 株式会社 OSA, 2026 年 11 月 16 日. https://gigazine.net/news/20261116-qwen4-max/

[6] **Artificial Analysis,** 「Qwen4 Max：知能・性能・料金の分析 (Intelligence, Performance & Price Analysis)」, Artificial Analysis Model Leaderboard (v5.0.1), 2026 年 11 月 16 日. https://artificialanalysis.ai/ja/models/qwen4-max

[7] Trending Topics, “Alibaba’s Giant Qwen4-Max Model Trails Claude, GPT and Moonshot AI”, Trending Topics EU, 2026 年 11 月 16 日. https://www.trendingtopics.eu/qwen4-max-alibaba-ranking-en/

[8] GIGAZINE, 「『Qwen4-Max-Preview』が無料で使用可能になったのでゲームを作ってみた」, 株式会社 OSA, 2026 年 11 月 12 日. https://gigazine.net/news/20261112-qwen4-preview/

[9] Fello AI, “Qwen 4 Release Date: Qwen4-Max Preview Is Out”, Fello AI News, 2026 年 11 月 11 日. https://felloai.com/qwen-4/

[10] Reddit Community (r/singularity), “Qwen4-Max: Your New Teammate Has Arrived (Community Benchmark & Evaluation Thread)”, Reddit, 2026 年 11 月 16 日. https://www.reddit.com/r/singularity/comments/1xxdj2b/qwen4max_your_new_teammate_has_arrived/

[11] Reddit Community (r/LocalLLaMA), “Observations from seeing Qwen 4 max's outputs / Regrets about Qwen4 branding and release strategy”, Reddit, 2026 年 11 月 11 日–12 日. https://www.reddit.com/r/LocalLLaMA/comments/1xtt3rb/observations_from_seeing_qwen_4_maxs_outputs/

[12] Reddit Community (r/LocalLLM), “Qwen officially announces Qwen4-Max and Qwen4 smaller models!”, Reddit, 2026 年 11 月 16 日. https://www.reddit.com/r/LocalLLM/comments/1xxe89d/qwen_officially_announces_qwen4max_and_qwen4/