Qwen4-Max 徹底調査レポート

2026 年 11 月 16 日の正式リリース ― 発表内容、性能評価、評判、日本企業にとっての留意点

調査・作成時点：2026 年 11 月 16 日（日本時間）

Claude Fable 5.2

本文書について

本文書は、2026 年 11 月 16 日に実施した調査の結果を、本文中の上付き番号と文末の参考文献リストで出典を示す形式に再構成したものである。記述は「確認された事実（一次ソースで裏付けあり）」「二次情報・報道ベース」「分析・推論」に区分した。

参考文献には、本文書の作成時に内容を再確認できたかどうかを【確認済】【一部確認】【未再確認】で示した。【未再確認】は、調査時には参照されたが本文書の作成時には再取得できなかった URL で、題名の一部は URL の表記から補っている。出典番号がなく「調査時の記録」と注記した記述は、対応する出典 URL を特定し直せなかったものである。

再確認の過程で判明した追補・修正は次の 4 点である。

(1) バッチ推論は、正式版ドキュメント（英語版、11 月 16 日更新）では北京リージョンのみ対応と記載されている[1]。調査時は「全リージョン非対応」と「北京のみ対応」の情報が併存していた。

(2) インプレスの PC Watch が、11 月 11 日にプレビュー版の公開を報じていた[2]。調査時に「Impress の報道は未確認」とした点は、プレビュー版については当たらない。

(3) JobBench の値は、調査時の記録（67.3）と DataLearner の記載（68.5）が食い違う[3]。

(4) 公式ブログは題名と冒頭の要旨までしか再取得できず、本文中の数値は報道との照合にとどまる[4]。

1. 要点

- **前提の真偽：**正しいと判断する。Qwen 公式 X[5]、公式ブログ[4]、Alibaba Cloud Model Studio の公式ドキュメント[1]により、北京時間 11 月 16 日午前に Qwen4-Max の正式版（モ

デル ID qwen4-max、従量課金の API）が公開されたことを確認した。ただし、公開されたのは Qwen4 ファミリーのうち Max のみで、Max 本体の重みは公開されていない[3,5]。

- **中身：**主幹 2.4 兆パラメータの MoE（トークンあたりの活性パラメータ 710 億）に、3,280 億の N-gram 埋め込みを加えた構成である[4]。コンテキストは 100 万トークンで、テキスト・画像・動画を入力でき、出力はテキストである[1]。国際版の料金は 100 万トークンあたり入力 1.5 ドル、出力 4.5 ドル、キャッシュ読み取り 0.075 ドルで[6,7]、前世代の Qwen3.8-Max（2 ドル／6 ドル）より 25%安い[8,9,10]。
- **評価：**Artificial Analysis の総合指数は 43 点である。Qwen3.8-Max（34 点）より 9 点高いが、Kimi K3.1 Pro（46 点）や Claude Opus 5.5 Max（59 点）には届かない[11,12,13]。Vals Index と Arena では、中国モデルの最上位と評価された[12,14]。公式の比較表は自社に有利な項目が目立ち、Terminal-Bench 4.0 などエージェント型コーディングの弱さが指摘されている[12,15,16]。
- **位置づけ（分析）：**絶対性能で首位を狙うモデルではなく、「価格性能比に優れた中国最上位クラスのクローズドモデル」とみるのが妥当である。日本企業が使う場合は、推論の実行地域と、自社環境に導入できない点が主な検討事項になる[1]。

2. 主な確認事項

2.1 確認された事実（一次ソースで裏付けあり）

- **発表主体と形態：**Qwen 公式 X は 11 月 16 日に正式版を告知し、API が QwenCloud で利用できること、Qwen Studio・QwenWork・Qoder でも使えること、小型の Qwen4 モデルの重みを今後公開することを明らかにした[5]。開発者向けアカウント（@QwenDevs）も同日に「Qwen4-Max is officially here」と投稿した[17]。公式ブログ「Qwen4-Max: Your New Teammate Has Arrived」も同日に公開されている[4]。
- **モデル ID と仕様：**モデル ID は qwen4-max。入力はテキスト・画像・動画、出力はテキスト。コンテキスト長は 1,000,000 トークン、非思考モードの最大出力は 131,072 トークン、思考と最終回答の合計上限は 262,144 トークンである。推論強度は low／high／xhigh の 3 段階で、デフォルトは high。API 形式は OpenAI Chat Completions、OpenAI Responses、

Anthropic Messages、DashScope の 4 種と互換がある。深い思考、並列ツール呼び出し、Web 検索、Function Calling、構造化出力、明示・暗黙キャッシュに対応する。ファインチューニングとモデルデプロイは全リージョンで非対応である[1]。

- **提供リージョン：**北京、シンガポール、フランクフルト、バージニア、東京、香港の 6 リージョンが記載されている。デプロイ範囲は、北京が「Chinese mainland」、シンガポールが「International」、東京を含む残りの 4 リージョンが「Global」と表記されている。バッチ推論は北京のみ対応である[1]。個人向けの Token Plan（Personal Edition）はシンガポールのみで提供され、デプロイモードは「Global」で、推論はグローバルに実行されると明記されている[18]。
- **料金：**100 万トークンあたり、入力 1.5 ドル、出力 4.5 ドル、暗黙キャッシュ入力 0.075 ドルである[6,7]。明示キャッシュの作成は 5 分で 1.875 ドル、1 時間で 3 ドル[7]。TPM 上限は 500 万である[19]。公式 X は、全項目で Qwen3.8-Max より安く、暗黙キャッシュは 70%値下げしたと説明している[5]。
- **前世代とアーキテクチャの系譜：**Qwen3.8-Flash-Next（2026 年 8 月 26 日公開のオープンウェイトモデル）は、公式モデルカードで、Qwen4 を支えるアーキテクチャの実験的プレビューと位置づけられていた[20,21]。総パラメータ 1,250 億（活性 60 億）で、Gated DeltaNet と Qwen Sparse Attention（QSA）のハイブリッド、Gated Residual、N-gram 埋め込みを採用する[21]。N-gram 埋め込みが 510 億パラメータであること、ライセンスが Qwen Community License 1.0 であることは、調査時にモデルカードと GitHub で確認した[20,22]。
- **前世代 Max の重み公開：**Alibaba が米 SEC に提出した 6-K（2026 年 6 月期決算）は、8 月に旗艦モデル Qwen3.8-Max を前版から 3 カ月以内に投入し、2.4 兆パラメータのモデル重みを公開したと記している[23]。Max シリーズの重みが公開されたのはこのときが初めてで[24]、今回の Qwen4-Max は再びクローズドに戻った。

2.2 二次情報・報道ベース

- Reuters は 11 月 15 日付で、Alibaba が、複雑な専門業務と科学研究に強く安全対策も強化した新しい旗艦モデルとして Qwen4-Max を発表したと報じた[25]。

- SCMP（Alibaba が所有する媒体）は、次世代の Qwen 5 を 5 兆〜10 兆パラメータに拡大する計画を報じた[26,27]。これはロードマップ上の目標であり、確定した仕様ではない。9 月の雲栖大会で示された、Qwen 4.5／Qwen 5 世代に向けた見通しとされる[10]。
- Hacker News では、公式ブログへのリンク投稿がトップページに入った（242 ポイント、82 コメント）[16]。

2.3 分析・推論

- 今回の発表の重点は、絶対性能で首位を取ることではなく、エージェント業務の単価を下げることにあるとみられる。活性パラメータの削減（950 億→710 億）、トークン消費の削減、単価の引き下げがそろっており、長時間のエージェント処理を安く回すための設計と読める。

3. 前提の真偽検証と発表の経緯

表 1　発表までの経緯（日付は北京時間）

日付	出来事	確度・出典
2026 年 8 月 26 日	Qwen3.8-Flash-Next（Qwen4 アーキテクチャのプレビュー）をオープンウェイトで公開	一次[20,21,22]
9 月 22 日	雲栖大会（Apsara）で Qwen4 が学習中であることを示し、Max／Plus／Flash／27B の 4 つの名称を提示（仕様と日付は示されず）	二次[10,28]
11 月 6 日	公式 X が「The becoming begins.」という予告を投稿	二次[29]。投稿時刻（20:00）と中国語の文言は調査時の記録
11 月 7 日	Qoder（Alibaba のコーディング IDE）に、正体を伏せたモデル「Qoduck」が登場	二次（調査時の記録）。Qwen4-Max の初期版という見方は未確認
11 月 11 日 10:00	「Qwen4 is coming.」の投稿とともに、qwen4-max-preview を Token Plan 限定で公開（11 月 25 日 24:00 に終了予定）	一次[30,31]。ITmedia[32]と PC Watch[2]が同日に報道
11 月 16 日 午前	**正式版 qwen4-max を公開。API、Token Plan、Qwen Studio、QwenWork、Qoder で提供**	一次[1,4,5]

- **時差の整理：**公式投稿は 02:01 UTC と記録されている[8]。北京時間 10:01、日本時間 11:01、米東部時間では 11 月 15 日 21:01 にあたる。Reuters が「日曜（11 月 15 日）」とし[25]、中国と日本の媒体が「11 月 16 日」とするのは時差によるもので、矛盾ではない。IT 之家の第一報は北京時間 10:07 である[6]。
- **名称：**正式名称は「Qwen4-Max」、モデル ID は qwen4-max である[1]。プレビュー版の ID は qwen4-max-preview で[31]、別名 qwen4.0-max-preview も使えるとされる（調査時の記録）。中国語では「千问 Qwen4-Max」「通义千问 Qwen4-Max」と表記される[6]。
- **「Qwen4」全体のリリースではない：**公開されたのはシリーズ最初のモデルである Max のみである。中小サイズのモデルは順次オープンソース化するとされるが、時期と型番は公表されていない[6,33]。

4. 発表内容（モデルの中身）

4.1 位置づけ

Qwen4 世代の旗艦にあたるマルチモーダル推論モデルである[1]。公式は、仕事・創作・発見のための「AI チームメイト」と表現し、3D モデリング、ゲーム開発、製造、金融、法律、医療、先端科学などの長時間の専門タスクを対象に掲げている[4,5]。

4.2 アーキテクチャ

- 疎な MoE 構造で、主幹は 2.4 兆パラメータ、トークンあたりの活性パラメータは 710 億である。これとは別に、3,280 億パラメータの N-gram 埋め込みを持つ[3,4,34]。活性パラメータは、前世代 Qwen3.8-Max の約 950 億[10]から約 25%減った。一部の暗号資産系ニュースは「活性 7,100 億」と記しているが、誤記と判断する[35]。
- 4 層のうち 3 層が Gated DeltaNet で、履歴を固定長の状態に圧縮する。残りの 1 層は QSA で、ブロック単位で関連する文脈を選ぶ。長い系列での注意計算のコストを抑える狙いである[34]。
- Gated Residual は、残差流を 4 本の並列ブランチに広げる仕組みである[34]。学習中に loss スパイクが一度も起きなかったという説明と、事前学習済みベースモデルの BPB（bits per

byte）が評価対象の 5 チェックポイント中で最も低く、前世代より 5〜13%改善したという説明は、調査時に報道で確認したもので、本文書の作成時には再確認できていない。

- エージェント向けに「Live Delivery」を導入した。まず使える初版を渡し、フィードバックに応じて版を重ねる方式で、Qwen-Live-Delivery-Plugin があわせてオープンソースで公開された[34]。

4.3 提供形態

- API は、QwenCloud（国際）と千問 AI プラットフォーム（百煉、Model Studio）で提供される。アプリでは、Qwen Studio、QwenWork（千問办公）、Qoder、千問アプリで使える[5,33]。
- 重みは公開されていない[3]。Artificial Analysis も「proprietary」に分類している[11]。Qwen3.8-Max では重みが公開されたため[23]、今回は方針の巻き戻しにあたる。

4.4 料金

表 2　料金の比較（100 万トークンあたり）

モデル	入力	出力	キャッシュヒット
Qwen4-Max（国際）[6,7]	1.50 ドル	4.50 ドル	0.075 ドル
Qwen4-Max（中国国内）[3,12,33]	10 元	30 元	0.5 元
Qwen3.8-Max（前世代、国際）[8,9,10]	2.00 ドル	6.00 ドル	0.25 ドル
Qwen3.8-Max（前世代、中国国内）[12]	12 元	36 元	1.5 元
Claude Opus 5.5（Anthropic 標準）[15]	4.00 ドル	20.00 ドル	0.20 ドル

- thinking トークンは出力単価で課金され、入力長による段階料金はない[36]。
- **1 タスクあたり費用の数字には揺れがある。**公式 X は「60%超の削減」[5]、IT 之家は Work360 で 26.58 元→11.16 元（58%減）[6]、AlphaSignal と Kingy AI は 4.43 ドル→1.67 ドル（約 62%減）[9,15]と伝えている。いずれも Alibaba の社内ベンチマークに基づく数字である。

4.5 前世代との違い

- 活性パラメータが約 25%減った。
- 構成が、GDN＋QSA＋N-gram 埋め込み＋Gated Residual に変わった[34]。
- 視覚がネイティブに統合された[4]。
- トークン消費が減った。Work360 では Xhigh で約 24%、High で約 35%、Terminal-Bench 4.0 では約 48〜49%の削減である[6]。Terminal-Bench 4.0 の Xhigh では、1 試行あたりの平均が 289K トークンから 151K トークンに減り、得点は 37.9 から 48.5 に上がった[37]。
- 単価が下がった（表 2）。

5. 性能・ベンチマーク

5.1 公式発表の値（自己申告。測定条件は完全には開示されていない）

- Work360（Alibaba 独自のベンチマーク。26 領域、182 サブ領域、721 の専門方向）：52.69 点。前世代は 47.74 点[34]。
- Agents' Last Exam：46.7 点[3]（前世代比+18.8 点）。Chartography：54.2 点（+12.2 点）。JobBench：67.3 点（GPT-6.1 Sol より上、Claude Opus 5.5 より下）。前世代比の差分、Chartography、JobBench の値は調査時の記録である。JobBench は、DataLearner では 68.5（Opus 5.5 は 75.4）とされており[3]、値は確定していない。
- セキュアコード（Last Secure Code Bench）：機能と安全性の両方を満たす関数レベルの合格率が 22.0%から 31.8%に上がり、関数・ファイル・リポジトリの 3 つの粒度すべてで安全に生成できた割合が 8%から 14.7%に上がった（調査時の記録）。

表 3　Alibaba が作成した 16 パネルの比較表からの抜粋

ベンチマーク	Qwen4-Max	Claude Opus 5.5
DeepSWE v1.1	75.7	74.2
Terminal-Bench 4.0	48.5	59.6
ProgramBench	31.5	65.0
Agents' Last Exam v1.1	46.7	43.4

ベンチマーク	Qwen4-Max	Claude Opus 5.5
HLAB-Vals 1.1（法務エージェント）	21.3	9.2
AutomationBench-AA	72.5	69.5
AgentArena net improvement	7.5	18.2
BioMysteryBench	85.0	81.9

出典：Kingy AI による転記[15]。AutomationBench-AA と BioMysteryBench の Opus 5.5 の値を除き、DataLearner の記載と一致する[3]。

- 16 パネルのうち、Qwen が上回るのは 6、Opus が上回るのは 10 である[15]。報道の見出しにある「Opus 超え」は、DeepSWE など一部の項目に限った話である。
- **表の中の数値にも食い違いがある。**Alibaba の表では Opus 5.5 の Terminal-Bench 4.0 が 59.6 だが、Vals は 65.15 としており[14]、Anthropic 自身は 66.4 と公表している（後者は調査時の記録）。ProgramBench の Opus の値も、Alibaba の表では 65.0、Vals では 18.50 である[14]。どちらが正しいかは本文書では判断できない。

5.2 第三者評価

表 4　第三者機関による評価

評価機関・指標	Qwen4-Max	比較対象	備考・出典
Artificial Analysis Intelligence Index v5.0.1	43（xhigh）	Kimi K3.1 Pro 46、Opus 5.5 Max 59、Qwen3.8 Max 34	1 タスクあたりのコストは Qwen 3.27 ドル、Opus Max 10.04 ドル[11,12,13]
Vals Index v2.2	63.20%（総合 11 位、中国モデル 1 位）	Opus 5.5 max 70.14%	1 テストあたりのコストは Qwen 9.33 ドル、Opus 24.77 ドル[12,14]
Vals Terminal-Bench 4.0	48.48%	Opus 5.5 65.15%	前世代は 38.9%[14]
Arena Code Arena WebDev	1,750（6 位、暫定）	Claude Fable 5.2 1,826、Opus 5.5 1,816	前世代 Qwen3.8-Max（0902）は 1,676[12,38]
Arena Agent Arena	14 位（net improvement +7.5%）	Opus 5.5 Max +18.19%	Bash 信頼性は 29 位[12,38]

- Vals は、Qwen 4 Max が Vals Index で中国モデルのトップ（総合 11 位）になったと投稿した[14]。Arena の数値はプレビュー期間を含むデータで、提供されたチェックポイントが正式版と同じかどうかは確認されていない。
- Artificial Analysis の同価格帯 81 モデルの中では 38 位で、中央値の 41 点をわずかに上回る[39]。
- Hacker News では、GPT-6.1 Sol（medium）が 46 点を平均 0.36 ドルで出している点を挙げ、Qwen4-Max のコスト優位は相対的なものだとする指摘があった[16]。
- Kingy AI が 18 題 96 回答の小規模テストを行ったところ、両モデルとも満点だった。課金額は Qwen の方が 78%低く、回答完了までの中央値は Opus の 2.76 秒に対して Qwen は 4.20 秒だった。設問が易しく、性能差は測れていない[15]。

5.3 公式主張と第三者評価の乖離（分析）

- 公式の表には、自社ベンチマーク（Work360）や、自社に有利な専門ドメイン（法務、生物、医療）が多く含まれる。総合指数（AA、Vals）や、エージェント型のターミナル操作・ソフトウェア工学（Terminal-Bench、ProgramBench、Agent Arena）では、Opus 5.5 との差がはっきり出ている。
- ただし、HLAB-Vals のように、第三者評価でも Qwen が Opus を大きく上回った項目がある。「すべてベンチマーク向けの最適化だ」と切り捨てるのは不正確である。強みは「ドキュメント・ツール型の専門業務」、弱みは「長いターミナル操作」と整理するのが妥当と考える。

5.4 同時期の競合

- **米国：**Anthropic（Claude Fable 5.2、Mythos 5.2、Opus 5.5）、OpenAI（GPT-6.1 Sol、GPT-6.1 Terra、GPT-6 Astra）、Google（Gemini 4 Carbon、Gemini 4 Hydrogen）。
- **中国：**Moonshot（Kimi K3.1 Pro。重みの無償公開を予告済み）、DeepSeek（V4.1 Flash）、Zhipu（GLM-5.3）、Xiaomi（MiMo-V2.6-Pro）。

- 以上のモデル名は、GIGAZINE、Kingy AI、Hacker News などに現れる名称によった[15,16,38]。GIGAZINE は、AA 指数で Claude Fable 5.2 や Gemini 4 Carbon とは大差がある一方、Claude Haiku 5.5 などコスト重視の米国モデルよりは高いと整理している[38]。
- xAI と Meta の最新モデルとの直接比較は確認できなかった（未確認）。

6. 評価・評判

6.1 中国国内（肯定的な声が中心で、細かい課題も指摘されている）

- 中国メディアは総じて、「2.4 兆パラメータ」「単タスク費用の約 6 割減」「Code Arena で世界 6 位、国産 1 位」といった点を前面に出して報じた[12,33,34,37]。
- 智東西の実測は、プレビュー版で見られたマルチモーダル生成の不一致（口の動きのずれなど）が正式版で改善したと評価した。一方、プレビュー段階の体験では、複雑なタスクで物理法則に反する出力が出るとの指摘もあった。網易号への転載記事[40,41]とみられるが、媒体との対応は再確認できていない。
- 知乎の実測記事では、3D、フロントエンド、ツール利用は「第一梯隊」との評価がある。ほかに、トークン消費は控えめで手戻りが少ないがプレビュー版の生成速度は約 33 tokens/s と遅い、長いタスクが途中で切れる、長尺アニメーションではフレームの一貫性が弱い、といった声があった[42]。
- 什么值得买の集約記事は、プレビュー版について、何度修正しても正常に動かない、テストコードを挿入する、事前に定めたルールに反する、といった利用者の報告を紹介している。正式版で直ったかどうかは公式発表からは分からないとしている[12]。
- X 上の GameFeel RE ランキングでプレビュー版が 10.3 点と Qwen3.8 系より低いという報告や、小紅書の否定的な投稿（確認できたのは題名のみ）もある。いずれも調査時の記録である。
- 微博、V2EX での体系的な反応は確認できなかった（未確認）。

6.2 英語圏

- Reuters は安全対策の強化に焦点を当てた。学習時にコードインジェクション、メモリ安全性、アクセス制御に関するチェックを組み込んだこと、サイバー犯罪や危険物質に関する悪用対策を強めたこと、社内テストで重大な高リスク能力の証拠はないとしたことを伝えている[25]。
- SCMP は、攻撃的なサイバー能力に関するデータを学習から除いたこと、CBRN（化学・生物・放射性物質・核）向けの専用の安全学習を行ったことを報じた[26]。SCMP は Alibaba 傘下であり、利益相反に留意が必要である。
- TechNode と Unite.AI は、価格の引き下げと正式提供の開始を中心に伝えた[43,44]。
- Hacker News の反応は割れた。AA 指数が振るわず Kimi K3.1 も超えていないという批判に対し、AA 指数は最近意味が薄れているという反論が出た。Terminal-Bench の低さを大きな欠点とする指摘、Qwen 3.8 は約 30 tok/s で Opus の約 80 tok/s より遅く第三者の推論提供者も少ないという速度面の懸念、オープンウェイトではないことの確認、本命は 27B と Flash のオープン版だという期待が目立った[16]。
- r/LocalLLaMA の正式版スレッドの内容は確認できなかった（未確認）。

6.3 日本語圏

- GIGAZINE（11 月 16 日 13 時 27 分）は、一部のテストで Claude Opus 5.5 を上回ったこと、小型のオープンモデルが後日公開予定であることを見出しに取り、AA 指数 43 点、Agent Arena 14 位、WebDev 6 位、価格を紹介した[38]。11 月 12 日にはプレビュー版でゲームを作る試用記事を出し、思考過程が中国語で表示されたこと、1 本作るのに約 20 分かかったことを記している[45]。
- ITmedia AI＋[32]と PC Watch[2]は、11 月 11 日のプレビュー公開を報じた。PC Watch は、同日時点でパラメータ数、ベンチマーク、正式提供の時期が公表されていないことを明記している[2]。
- note（CHINA CASE）は、自動車製造の加工工数が 95%減ったこと、AI 冷却液の新しい分子を発見したことを紹介している（会員限定記事）[46]。

- 日経の報道は確認できなかった（未確認）。インプレスについても、正式版に関する記事は確認できていない。

6.4 市場の反応

- Alibaba の香港株（9988）は、11 月 16 日の午前終了時点で 1.05%高の 91.50 香港ドルだった[47]。前営業日の 13 日は 3.46%安の 90.55 香港ドルで引けている（調査時の記録）。
- 米国 ADR（BABA）は、日曜夜の時間外取引で 1%超下げた[48]。16 日の終値は確認できていない。
- 同じ日に、Morgan Stanley が、GPU 不足により今後数年アリババのクラウド収入は圧迫されると指摘したと報じられた（調査時の記録）。Qwen4-Max を名指ししたアナリスト評価は見つかっていない。決算発表は 11 月 19 日の予定とされる。

7. 今後の影響・意味合い（分析・推論）

- **米中の差：**AA 指数では、米国の旗艦（Opus 5.5 Max 59 点）と Qwen4-Max（43 点）の間に 16 点の開きがあり、中国勢の中でも Moonshot が先行している[11]。Hacker News では、オープン系との差は広がっているのではないかという問いも出た[16]。今回のデータだけでは「何カ月遅れか」を数値で示すことはできない。総合性能で差を縮めたというより、価格性能比と特定の専門業務で戦う路線が鮮明になった、と読むのが妥当である。
- **中国国内の競争：**Moonshot（Kimi K3.1 Pro）が総合指数で先行し、重みも公開している。これに対して Alibaba は、旗艦をクローズドに戻し、小型モデルのオープン化で開発者のエコシステムを維持する二本立ての戦略をとる[5]。Hacker News では、Zhipu の GLM 次期版にも関心が集まった[16]。
- **計算資源の制約：**活性パラメータの削減と N-gram 埋め込みは、メモリが限られたアクセラレータでも効率よくパラメータを増やすための設計と説明されており[20]、GPU 調達の制約に合わせたものと読める。Alibaba の 4〜6 月期の設備投資は 677 億元（99.75 億ドル）で、前年同期の 386.76 億元から 75%増えた[23]。投資は続いているが、GPU 不足の指摘を踏まえると、提供速度や混雑時のキャパシティが今後のリスクになる。
- **日本企業の留意点：**

(1) Token Plan はシンガポール限定で、推論は Global である（国外で処理される可能性がある）[18]。

(2) API の対応リージョンに東京が含まれるが、公式ドキュメントでは東京のデプロイ範囲が「Global」と表記されている[1]。東京で受けたリクエストが日本国内で推論されるかどうかは、契約前に確認が必要である。

(3) 自社環境へのデプロイとファインチューニングはできない[1]。

(4) 中国企業のサービスであることに伴う規制、経済安全保障、データ取扱い（Model Studio のデータ保持ポリシー、学習データの開示）を個別に審査する必要がある。

- **知財業務への示唆：**

(1) 100 万トークンのコンテキストは、明細書や引用文献をまとめて読み込ませる用途に向く[1]。

(2) 法務エージェント評価の HLAB-Vals では Opus 5.5 を大きく上回った[3,15]が、Legal Research Bench（Qwen 41.35%、Opus 50.48%）では下回っている（調査時の記録）。法的な調査や判断をそのまま任せられる水準ではない。

(3) 特許翻訳や先行技術調査に特化した評価は確認できなかった。

(4) 出願前の発明情報は機密性が高く、国外で推論される API への入力は避けるべきである。将来公開される小型のオープンウェイト版 Qwen4 を、社内環境で検証するのが現実的である。

8. 提言

1. **本番移行の前に評価から始める。**大量の要約・抽出・下書き作業や、ドキュメントとツールを多用するエージェント業務で、Claude Opus 5.5 や社内の標準モデルと並べて比較する。トークン単価ではなく、「採用できた成果 1 件あたりの総費用」で判断する。
2. **エージェント型コーディングには慎重になる。**Terminal-Bench、ProgramBench、Agent Arena、Bash 信頼性のいずれも弱い。リポジトリ規模の自律開発は米国の旗艦モデルを優先し、Qwen4-Max は補助的な位置に置くのが無難である。

3. **データの所在を確認する。**東京リージョンのデプロイ範囲、データ保持、学習への利用の有無を契約書面で確認するまで、機密データ（出願前の技術情報、個人情報）を入れない。
4. **プレビュー版の ID から移行する。**qwen4-max-preview は 11 月 25 日 24:00（UTC+8）で終了する[31]。正式版の qwen4-max に切り替え、出力上限が 393,216 から 262,144 に変わる点にも対応する[1,31]。
5. **オープンウェイトの小型版を待つ価値がある。**自社環境での運用が必要な知財業務や機密業務では、Qwen4 の 27B や Flash のオープン版が出た段階で、ライセンス（Apache 2.0 か Qwen Community License か）を確認したうえで評価するのが現実的である。

9. 留意事項

- **リリース当日の情報である。**Arena の値は「暫定」扱いで、プレビュー版を含むデータの可能性がある。AA、Vals、Arena の値は今後変わりうる。
- **一次ソースへのアクセスに制約がある。**Qwen 公式ブログの本文と SCMP の記事本文は取得できなかった[4,26]。数値の多くは、複数の報道が一致していることで確認している。
- **数値の不一致がある。**1 タスクあたりの費用削減率（58%、60%超、62%）は、通貨や計算方法の違いで揺れている。最大出力は、正式版ドキュメントの 262,144[1]と、QwenCloud のマーケットプレイス表記や DataLearner の 393K[3,19]が食い違う。JobBench の値（67.3／68.5）も食い違う[3]。
- **信頼性の低い情報源は事実認定に用いていない。**暗号資産系のニュース速報（KuCoin など）、SEO 目的とみられる比較サイト、出所不明の「AA 指数 53 点」チャートがこれにあたる。53 点のチャートについては、AA の公式ランキングには存在しないとの指摘がある[9]。Wikipedia の記述には「要出典」の箇所がある[49]。什么值得买の記事は AI による集約記事である[12]。
- **前世代 Max の公開状況をめぐり、情報源が相反する。**Alibaba の 6-K は Qwen3.8-Max の重み公開を明記している[23]。一方、NYU 上海の記事には「2025 年 9 月以降の Max はすべて API 限定」とするものがある[50]。同サイトは 8 月の重み公開を伝える記事も掲載しており

[24]、記事の時点の違いによる可能性がある（推論）。本文書では一次ソースである 6-K を優先した。

- **未確認の事項：**9988 の 16 日終値、Qwen4-Max を名指ししたアナリスト評価、日経の報道、r/LocalLLaMA と微博の反応、xAI・Meta の最新モデルとの比較、Eddie Wu CEO の発言が直接の引用かどうか。

参考文献

URL の閲覧日はいずれも 2026 年 11 月 16 日である。各項目末尾の表示は、本文書の作成時に内容を再確認できたかどうかを示す。【確認済】は題名と該当する記述を確認できたもの、【一部確認】は一部のみ確認できたもの、【未再確認】は調査時には参照されたが再取得できなかったものである。

[1] Alibaba Cloud「Qwen4-Max」Alibaba Cloud Model Studio ドキュメント（英語版）、最終更新 2026 年 11 月 16 日。 https://www.alibabacloud.com/help/en/model-studio/qwen4-max **【確認済】**調査時は中国語版（http://docs.modelstudio.console.alibabacloud.com/zh/model-studio/qwen4-max）を参照した。

[2] 竹元かつみ「「Qwen4」早くも登場、最上位「Qwen4-Max」を無料でも試せる」PC Watch（インプレス）、2026 年 11 月 11 日 12:03。 https://pc.watch.impress.co.jp/docs/news/2155430.html **【確認済】**本文書の作成時に追加した出典。

[3] DataLearner「Qwen4-Max 正式发布：2.4T/71B MoE、16 项官方基准与最新 API 定价」、2026 年 11 月 16 日更新。 https://www.datalearner.com/ai-models/pretrained-models/qwen4-max **【確認済】**公式資料を二次的に集約したサイト。

[4] Qwen「Qwen4-Max: Your New Teammate Has Arrived」Qwen 公式ブログ、2026 年 11 月 16 日（北京時間）。 https://qwen.ai/blog?id=qwen4-max **【一部確認】**題名と冒頭の要旨のみ確認。本文は取得できなかった。

[5] Qwen（@Alibaba_Qwen）、X への投稿（Qwen4-Max 正式版の告知）、2026 年 11 月 16 日。 https://x.com/Alibaba_Qwen/status/2122148127520289151 **【確認済】**

[6] 问舟「阿里通义千问 Qwen4-Max 正式上线：任务成本降低超 60%，后续将开源小尺寸模型」IT 之家、2026 年 11 月 16 日 10:07。 https://www.ithome.com/1/027/512.htm **【確認済】**調査時は新浪科技などの転載で参照。原記事の URL は本文書の作成時に追加した。

[7] OpenRouter「Qwen: Qwen4 Max – API Pricing & Providers」、2026 年 11 月 16 日閲覧。 https://openrouter.ai/qwen/qwen4-max **【確認済】**本文書の作成時に追加した出典。

[8] AI/TLDR「Qwen4-Max」リリースページ。 https://ai-tldr.dev/releases/qwen-qwen4-max/ **【未再確認】**

[9] AlphaSignal「Alibaba's Qwen4-Max slashes agent task costs 62% with 1M-token context」。 https://alphasignal.ai/news/alibaba-s-qwen4-max-slashes-agent-task-costs-62-with-1m-token-context **【未再確認】**題名は URL 表記による。

[10] Alistair Wren「Qwen 4 Max 在 Apsara 2026 上发布：Alibaba 确认了什么，又没有确认什么」OrcaRouter ブログ（中国語版）。 https://www.orcarouter.ai/zh-CN/blog/qwen-4-max-lineup-announced-apsara-2026 **【確認済】**API ルーター事業者のブログ。同記事の各言語版で内容を確認した。本文書の作成時に追加した出典。

[11] Artificial Analysis「Qwen4-Max」モデルページ。 https://artificialanalysis.ai/models/qwen4-max **【未再確認】**

[12] 「通义千问 Qwen4-Max 正式上线：降本 60%的账怎么算、谁最受益」什么值得买、2026 年 11 月 16 日 14:10。 https://post.smzdm.com/p/aomqg766/ **【確認済】**複数の投稿を AI で集約した記事（AIGC）と明記されており、信頼性は限定的。

[13] Artificial Analysis「Alibaba Models: Intelligence, Performance & Price」。 https://artificialanalysis.ai/models/creators/alibaba **【確認済】**Qwen3.8 Max（0902）の指数 34 を確認した。本文書の作成時に追加した出典。

[14] Vals AI（@ValsAI）、X への投稿、2026 年 11 月 16 日。 https://x.com/ValsAI/status/2122152965293457822 **【未再確認】**

[15] Kingy AI「Qwen4-Max vs Claude Opus 5.5」。 https://kingy.ai/blog/qwen4-max-vs-claude-opus-5-5/ **【未再確認】**題名は URL 表記による。

[16] Hacker News「Qwen4-Max: Your New Teammate Has Arrived」スレッド（item 50513215）。 https://news.ycombinator.com/item?id=50513215 **【未再確認】**

[17] Qwen Devs（@QwenDevs）、X への投稿、2026 年 11 月 16 日。 https://x.com/QwenDevs/status/2122156218928242799 **【未再確認】**

[18] Alibaba Cloud「Token Plan Personal Edition overview」Alibaba Cloud Model Studio ドキュメント。 https://www.alibabacloud.com/help/en/model-studio/token-plan-personal-overview **【未再確認】**題名は URL 表記による。

[19] QwenCloud「qwen4-max」モデルページ。 https://www.qwencloud.com/models/qwen4-max **【未再確認】**

[20] Qwen「Qwen3.8-Flash-Next」モデルカード、Hugging Face。 https://huggingface.co/Qwen/Qwen3.8-Flash-Next **【一部確認】**カードの記述は転載と二次記事を通じて確認した。

[21] 「Qwen3.8-Flash-Next Previews Qwen4 Architecture With Hybrid Attention and 6B Active Parameters」Unite.AI。 https://unite.ai/qwen3-8-flash-next-previews-qwen4-architecture-with-6b-active-parameters **【確認済】**本文書の作成時に追加した出典。

[22] QwenLM「Qwen3.8-Flash-Next」GitHub リポジトリ。 https://github.com/QwenLM/Qwen3.8-Flash-Next **【未再確認】**

[23] Alibaba Group Holding Limited「Form 6-K（2026 年 6 月期四半期決算、Exhibit 99.1）」米国証券取引委員会（SEC）EDGAR、2026 年 8 月 20 日。 https://www.sec.gov/Archives/edgar/data/0001577552/000110465926099220/tm2623667d1_ex99-1.htm **【確認済】**Qwen3.8-Max に関する記述を検索結果の抜粋で確認した。設備投資の伸び（677 億元、75%増）は複数の報道でも確認した。

[24] 「Qwen3.8-2.4T-A95B: Alibaba Open-Weights Its Max-Tier Flagship」NYU Shanghai RITS。 https://rits.shanghai.nyu.edu/ai/qwen3-8-2-4t-a95b-alibaba-open-weights-its-max-tier-flagship/ **【確認済】**本文書の作成時に追加した出典。

[25] Reuters「Alibaba unveils new flagship AI model with safety upgrades」（93.3 The Drive 掲載）、2026 年 11 月 15 日。 https://www.933thedrive.com/2026/11/15/alibaba-unveils-new-flagship-ai-model-with-safety-upgrades/ **【未再確認】**

[26] South China Morning Post「Alibaba says new flagship AI model cuts task costs 60% while tightening cyber guardrails」。 https://www.scmp.com/tech/big-tech/article/3374540/alibaba-says-new-flagship-ai-model-cuts-task-costs-60-while-tightening-cyber-guardrails **【未再確認】**題名は URL 表記による。調査時も本文は取得できていない。

[27] OrcaRouter ブログ「Qwen 4.5 and Qwen 5 roadmap: 5–10T parameters」。 https://www.orcarouter.ai/blog/qwen-4-5-and-qwen-5-roadmap-5-10t-parameters **【未再確認】**題名は URL 表記による。

[28] Alexey Fateev（@superalesha）、X への投稿（Qwen4 の事前情報の整理）、2026 年 11 月 9 日ごろ。 https://x.com/superalesha/status/2119705653103018065 **【確認済】**

[29] 「Qwen4-Max in preview: Token Plan from $3 for new users, trial until November 25」 pasqualepillitteri.it、2026 年 11 月 11 日。 https://pasqualepillitteri.it/en/news/31971/qwen4-max-preview-alibaba-token-plan-from-3-dollars **【確認済】**個人運営のニュースサイト。本文書の作成時に追加した出典。

[30] Qwen（@Alibaba_Qwen）、X への投稿「Qwen4 is coming.」、2026 年 11 月 11 日。 https://x.com/Alibaba_Qwen/status/2120335993643684315 **【確認済】**本文書の作成時に追加した出典。

[31] Alibaba Cloud「Latest model: Qwen4-Max-Preview」Alibaba Cloud Model Studio ドキュメント（英語版）、最終更新 2026 年 11 月 11 日。 https://www.alibabacloud.com/help/en/model-studio/qwen4-max-preview **【確認済】**本文書の作成時に追加した出典。

[32] ITmedia AI＋、Qwen4-Max プレビュー版の公開に関する記事、2026 年 11 月 11 日。 https://www.itmedia.co.jp/aiplus/article/2611/11/2000003098/ **【未再確認】**題名は未確認。

[33]「阿里正式发布 Qwen4，单任务平均费用大降 6 成」搜狐（界面新闻の報道を転載）、2026 年 11 月 16 日 11:03。 https://www.sohu.com/a/1102110880_122014422 **【確認済】**

[34]「2.4 万亿参数阿里通义千问 Qwen4-Max 正式上线：任务成本降低超 60%，后续将开源小尺寸模型」新浪科技（IT 之家配信）、2026 年 11 月 16 日 10:07。 https://finance.sina.com.cn/tech/digi/2026-11-16/doc-iniywwza8955380.shtml **【確認済】**

[35] KuCoin News「Qwen4-Max launches with 25% fewer active parameters and 62% lower task costs」。 https://www.kucoin.com/news/flash/qwen4-max-launches-with-25-fewer-active-parameters-and-62-lower-task-costs **【未再確認】**暗号資産系のニュース速報で信頼性は低い。題名は URL 表記による。誤記のある記事との対応は未再確認。

[36] SandBase ブログ「Qwen4-Max generally available」。 https://blog.sandbase.ai/qwen4-max-generally-available-sandbase-2026/ **【未再確認】**題名は URL 表記による。

[37] 網易科技、Qwen4-Max 正式版に関する記事、2026 年 11 月 16 日。 https://www.163.com/tech/article/LBROR49P00098IEO.html **【未再確認】**題名は未確認。

[38] GIGAZINE、Qwen4-Max 正式版に関する記事、2026 年 11 月 16 日 13:27。 https://gigazine.net/news/20261116-qwen4-max/ **【未再確認】**題名は未確認。

[39] Trending Topics、Qwen4-Max のランキングに関する記事。 https://www.trendingtopics.eu/qwen4-max-alibaba-ranking-en/ **【未再確認】**題名は未確認。

[40] 網易号掲載記事（媒体名・題名は未確認）、2026 年 11 月 16 日。 https://www.163.com/dy/article/LBRSI97O0556703U.html **【未再確認】**

[41] 網易号掲載記事（媒体名・題名は未確認）、2026 年 11 月 16 日。 https://www.163.com/dy/article/LBS22V02051180F7.html **【未再確認】**

[42] 知乎専欄、Qwen4-Max の実測記事（4 本）。 https://zhuanlan.zhihu.com/p/2103978083808186465 ／ https://zhuanlan.zhihu.com/p/2105289405908907348 ／ https://zhuanlan.zhihu.com/p/2104517472426648225 ／ https://zhuanlan.zhihu.com/p/2103905620697133459 **【未再確認】**

[43] TechNode「Alibaba launches Qwen4-Max with lower API prices and a million-token context window」、2026 年 11 月 16 日。 https://technode.com/2026/11/16/alibaba-launches-qwen4-max-with-lower-api-prices-and-a-million-token-context-window/ **【未再確認】**題名は URL 表記による。

[44] Unite.AI「Alibaba Cloud releases Qwen4-Max with 25% lower Singapore API prices」。 https://www.unite.ai/alibaba-cloud-releases-qwen4-max-with-25-lower-singapore-api-prices/ **【未再確認】**題名は URL 表記による。

[45] GIGAZINE（英語版）、Qwen4-Max プレビュー版の試用記事、2026 年 11 月 12 日。 https://gigazine.net/gsc_news/en/20261112-qwen4-preview/ **【未再確認】**題名は未確認。

[46] CHINA CASE（note）、Qwen4-Max の産業応用に関する記事（会員限定）。 https://note.com/china_case/n/n1ca57307d29a **【未再確認】**題名は未確認。

[47] 明報財經「恒指半日升 197 點 北水流入 4 億元 新世界跌 38…」、2026 年 11 月 16 日。 https://finance.mingpao.com/fin/instantf/20261116/1794801705502/%E6%81%92%E6%8C%87%E5%8D%8A%E6%97%A5%E5%8D%87197%E9%BB%9E-%E5%8C%97%E6%B0%B4%E6%B5%81%E5%85%A54%E5%84%84%E5%85%83-%E6%96%B0%E4%B8%96%E7%95%8C%E8%B7%8C38 **【未再確認】**題名は URL 表記による（末尾は省略）。

[48] Stocktwits「BABA Stock Edges Lower Overnight: Alibaba Reportedly Launches Most Advanced AI Model That Cuts Task Costs By 60%」。 https://stocktwits.com/news-articles/markets/equity/baba-stock-edges-lower-overnight-alibaba-reportedly-launches-most-advanced-ai-model-that-cuts-task-costs-by-60/cNb7gKGR9Xf **【未再確認】**題名は URL 表記による。

[49] Wikipedia（英語版）「Qwen」。 https://en.wikipedia.org/wiki/Qwen **【未再確認】**

[50] NYU Shanghai RITS「Qwen Restates Its Open-Weight Promise; Max Tier Still Closed」。 https://rits.shanghai.nyu.edu/ai/qwen-restates-its-open-weight-promise-max-tier-still-closed/ **【未再確認】**題名は URL 表記による。

付　調査時の出典一覧のうち、本文で番号を付していない URL

調査時に参照されたが、他の出典と内容が重複する、信頼性が低い、または本文の記述との対応を特定し直せなかったために、番号を付していないものである。いずれも本文書の作成時には再確認していない。

- 雪球「BABA」銘柄ページ　https://www.xueqiu.com/S/BABA
- 什么值得买 掲載記事　https://post.smzdm.com/p/a5rp0nw8/
- 什么值得买 掲載記事　https://post.smzdm.com/p/anvm5400/
- bannedbook.org 掲載記事　https://www.bannedbook.org/bnews/itnews/20261116/2381419.html
- KuCoin News（速報）　https://www.kucoin.com/news/flash/alibaba-launches-qwen4-max-ai-model-claims-coding-performance-edge
- KuCoin News（速報）　https://www.kucoin.com/news/flash/alibaba-launches-qwen4-max-with-50-lower-token-usage-and-28-higher-score
- Life News Agency　https://lifenewsagency.com/alibaba-launches-qwen4-max-flagship-ai-model/
- 網易号（IT 之家の転載）　https://www.163.com/dy/article/LBRN6GJK0511B8LM.html

- LavX News https://news.lavx.hu/article/alibaba-launches-qwen4-max-with-strict-cybersecurity-guardrails
- Crypto Briefing https://cryptobriefing.com/alibaba-unveils-qwen4-max-ai-model/
- NewsBytes https://newsbytesapp.com/news/science/alibaba-s-qwen4-max-claims-superior-performance-in-coding-and-cybersecurity-tasks/story
- Hacker News トップページ https://news.ycombinator.com/
- The News International（Reuters 配信） https://www.thenews.com.pk/latest/1423766-alibaba-unveils-flagship-ai-model-qwen4-max-with-safety-upgrades
- MLQ.ai https://mlq.ai/news/alibaba-launches-qwen4-max-saying-it-cuts-task-costs-by-about-60-with-tighter-cyber-guardrails/
- AI First Look https://aifirstlook.com/qwen-4/
- TechTimes（プレビュー版終了の報道） https://www.techtimes.com/articles/330536/20261112/alibaba-qwen4-max-preview-retires-november-25-no-benchmarks-model-card-price.htm
- X（@ItsmeAjayKV）の投稿 https://x.com/ItsmeAjayKV/status/2120249520835723623
- eesel AI ブログ https://www.eesel.ai/blog/qwen-4-max
- DoNews https://www.donews.com/news/detail/1/6772283.html
- Samaa TV（Reuters 配信） https://www.samaa.tv/2087360370-alibaba-unveils-new-flagship-ai-model-with-safety-upgrades
- Qwen 公式ブログ「Qwen3.8-Flash-Next」 https://qwen.ai/blog?amp=&amp=&amp=&id=qwen3.8-flash-next
- 騰訊新聞 https://news.qq.com/rain/a/20261116A05VFH00?id=20261116A05VFH00&path=a&app=news&suid=&redirect_pc=1
- Business Standard https://www.business-standard.com/world-news/alibaba-says-qwen4-max-cuts-task-costs-by-60-boosts-cyber-safeguards-126111600895_1.html
- ilSoftware https://www.ilsoftware.it/qwen4-max-la-nuova-ai-di-alibaba-sfida-claude-opus-5-5/
- X トレンドページ https://x.com/i/trending/2122164059252437215
- 新浪科技 掲載記事 https://finance.sina.com.cn/tech/roll/2026-11-16/doc-iniywwza8990839.shtml
- Asianet Newsable（Stocktwits 配信） https://newsable.asianetnews.com/markets/baba-stock-edges-lower-overnight-alibaba-reportedly-launches-most-advanced-ai-model-that-cuts-task-costs-by-60-articleshow-xzm1scz
- Kantan News https://kantan.news/news/alibaba-announces-qwen4-max-model-60-percent-reduction-in-processing-costs
- note（s_kashiwakura） https://note.com/s_kashiwakura/n/n9d13f234dd48
- Investing.com 香港 https://hk.investing.com/news/stock-market-news/article-1759104
- Benzinga https://www.benzinga.com/markets/tech/26/11/63174214/why-is-alibaba-stock-falling-thursday
- eesel AI ブログ（レビュー） https://www.eesel.ai/blog/qwen-4-max-review
- Tech Insider https://tech-insider.org/qwen4-four-variants-no-confirmed-open-weights-alibaba-2026/
- APIYI ヘルプ https://help.apiyi.com/en/qwen4-max-vs-claude-opus-5-5-coding-agent-pricing-en.html