ChatGPT

Alibaba Qwen3.8-Max 調査報告

調査基準日：2026年8月4日（日本時間）

エグゼクティブサマリー

Alibaba／Qwenチームが2026年8月3日に正式発表した **Qwen3.8-Max** は、総パラメータ数 **2.4兆、推論時の有効パラメータ数95B** のSparse Mixture-of-Expertsモデルである。Qwen 3.5系のアーキテクチャを基盤とし、Sparse MoEと「hybrid attention」を組み合わせ、テキスト・画像・動画を入力できるネイティブ・マルチモーダルモデルとして提供される。最大コンテキスト長は100万トークン、最大出力長は131,072トークンである。2026年8月4日時点ではAPI提供が先行し、モデル重みは「翌週」にHugging FaceとModelScopeで公開予定とされている。したがって、現時点のQwen3.8-Maxは、厳密には**オープンウェイトモデルではなく、将来のオープンウェイト化が予告されたAPIモデル**である。 [1]

性能面では、Alibabaが公開した比較表において、旧Qwen3.7-Maxからコーディング、ツール利用、専門業務、推論の多くで大きく改善している。特にPaperBench 93.0、IFBench 82.8、OSWorld-Verified 86.1などが強い。一方、SWE-bench Proでは67.7でFable 5の80.0を下回り、Terminal-Bench 2.1では86.6でGPT-5.6 Solの88.8を下回る。LongBench v2も66.3であり、100万トークンという最大長が、そのまま最高水準の長文検索・理解精度を意味するわけではない。これらは主としてAlibaba自身が選択・実行した評価であり、第三者による再現結果ではない。 [2]

第三者評価として最も明確なのはLMArenaである。2026年8月4日時点で、Text Arenaは5位・1496±10、Vision Arenaは2位・1305±9、WebDev Arenaは4位・1668±18で、価格は入力100万トークン当たり2ドル、出力6ドルと表示されている。ただし、投票数はそれぞれ3,327、5,038、1,563で、全項目が「Preliminary」である。上位のClaude系モデルより信頼区間が広く、順位は今後相当に変動し得る。 [3]

最大の弱点は、**技術的透明性と安全性情報の不足**である。2026年8月4日時点で、Qwen3.8-Max固有のモデルカード、技術白書、学習データ概要、トークナイザー仕様、レイヤー数、専門家数、ルーティング方式、学習計算量、学習用ハードウェア、訓練精度、量子化成果物、標準化されたレイテンシ、安全性・毒性・脱獄・サイバー・生物化学リスク評価は確認できない。MMLU、HELM、LLM-Score、HumanEvalの正式結果も発表資料には含まれていない。重みとライセンスが未公開であるため、「オープンソース」という表現についても、再現可能性や商用利用条件を現段階で判断すべきではない。 [4]

実務上の評価は、**低価格なフロンティア級エージェントモデルとして有望だが、全面移行には早い**、となる。コーディング支援、リポジトリ分析、長文書・動画解析、オフィス業務、視覚を伴うGUI操作の候補として優先的に検証する価値がある。一方、長時間エージェントではツール呼び出しが構造化されず失われるというプレビュー版の報告があり、金融・法務・医療・インフラ変更などでは、権限制限、サンドボックス、出力検証、人間承認、既存モデルへのフォールバックが必須である。 [5]

公開状況と一次資料の監査

一次資料の公開状況を整理すると、通常のフロンティアモデル公開に必要な資料のうち、製品・API情報は充実しているが、モデル内部と安全性に関する資料はほぼ未公開である。

資料区分	2026年8月4日時点の状況	評価
公式発表	Alibaba Cloudが2026年8月3日に正式発表	確認済み
Qwen公式技術ブログ	Qwen公式ブログとAlibaba Cloud版が公開	確認済み
API仕様・価格	Alibaba Cloud Model Studio、QwenCloudで公開	確認済み
オンラインデモ	QwenCloudの「Try AI」、QwenWorkへの導線あり	確認済み
モデルカード	Qwen3.8-Max固有のものは未確認	未公開
技術白書	Qwen3.8-Max固有のものは未確認	未公開
モデル重み	翌週にHugging Face／ModelScopeで公開予定	未公開
GitHubモデル実装	Qwen3.8-Max固有リポジトリは未確認	未公開
ライセンス	重み公開前で、Qwen3.8-Max固有条件は不明	未指定
安全性／システムカード	未確認	未公開
学習データ概要	未確認	未公開

公式発表は、2.4兆パラメータ、100万トークン、LMArena順位、API提供、翌週の重み公開を明示している。Qwen技術ブログには比較ベンチマーク、API例、`reasoning_effort`、OpenAI互換・Anthropic互換インターフェース、Claude CodeやCodexなどとの統合例が掲載されている。[6]

モデル重みについて、公式ブログは「Qwen-Maxクラスで初めてオープンソース化する」と説明しているが、同時に具体的な公開日は「next week」としている。ライセンス本文、再配布条件、商用利用条件、派生モデル条件、Acceptable Use Policyはまだ示されていない。過去のQwenモデルがApache 2.0で公開された例があっても、Qwen3.8-Maxに同じ条件が適用されるとは限らない。重み、設定ファイル、モデルカードに記載されるライセンスを確認するまでは「Apache 2.0予定」などと推定すべきではない。[7]

Qwen3 Technical Reportは存在するが、これは2025年に発表された0.6B〜235BのQwen3系列を扱う文献であり、2.4TのQwen3.8-Maxを記述した白書ではない。Qwen3.8-Maxの仕様をQwen3報告書から補完することは、レイヤー数、専門家構成、トークナイザー、学習方法などを誤って推定する原因になる。[8]

発表資料には、約16日間にわたって自律的にソフトウェア開発を行い、`oh-my-cli`というエージェントフレームワークを生成した事例も掲載されている。この成果物はGitHubで検査可能なデモ・アーティファクトとして有用だが、「モデルが人間の介入なしに一貫して16日稼働できる」という一般的な信頼性保証にはならない。タスク選定、初期プロンプト、実行環境、失敗試行、人的介入、API再試行、総トークン数、コストの完全な監査記録は公表資料からは分からない。[9]

技術仕様と実装上の含意

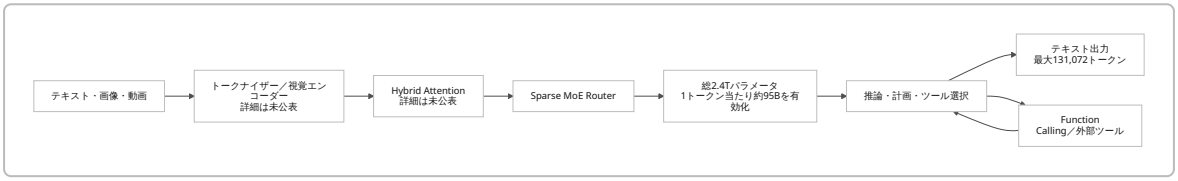
公開済みの仕様と未指定事項を明確に分ける必要がある。

項目	公開情報	確度・注意点
基盤	Qwen 3.5のアーキテクチャ基盤	公式明記
基本構造	Sparse Mixture-of-Experts＋hybrid attention	詳細方式は未指定

項目	公開情報	確度・注意点
総パラメータ数	2.4兆	公式明記
推論時有効パラメータ	95B	公式明記
入力モダリティ	テキスト、画像、動画	公式API仕様
出力モダリティ	テキスト	公式API仕様
コンテキスト長	1,000,000トークン	公式API仕様
通常時最大入力	991,808トークン	公式API仕様
思考モード最大入力	983,616トークン	公式API仕様
最大出力	131,072トークン	公式API仕様
最大思考チェーン	262,144トークン	公式API仕様
Function Calling	対応	公式API仕様
Structured Output	対応	公式API仕様
コンテキストキャッシュ	対応	公式API仕様
バッチ推論	非対応	Model Studio
APIファインチューニング	非対応	Model Studio
推論深度	`xhigh`、`medium`、`low`	`xhigh`が既定
API互換性	OpenAI Chat／Responses、Anthropic互換	完全な動作互換性を意味しない
トークナイザー／語彙数	未指定	モデルカード待ち
レイヤー数／隠れ次元	未指定	モデルカード待ち
専門家数／選択数	未指定	モデルカード待ち
位置表現	未指定	RoPE／YaRN等を推定不可
学習データ量・内訳	未指定	言語比率、コード、画像、動画、合成データとも不明
学習ハードウェア	未指定	GPU／Ascend等を推定不可
学習計算量	未指定	FLOPs、GPU時間とも不明
学習・推論精度	未指定	BF16、FP8等を推定不可
公式量子化	未指定	FP8、INT8、INT4、GGUF等は未公開
TTFT／tokens per second	未指定	数値ベンチマークなし

項目	公開情報	確度・注意点
知識カットオフ	未指定	Web検索機能とは別問題
Instruction tuning	実施内容未指定	SFTデータや手法は不明
RL／選好最適化	改善にRLを用いたとの記述はあるが詳細なし	アルゴリズム・データ・報酬モデル不明
LoRA／QLoRA	現時点で未提供	重み公開後も実装検証が必要

アーキテクチャについて公式に分かるのは、Qwen 3.5を基盤とするSparse MoE、hybrid attention、2.4T総パラメータ、95B有効パラメータまでである。通常のDense 2.4Tモデルより1トークン当たりの演算量を大幅に抑えられると考えられるが、95Bだけをメモリーに置けば動作するという意味ではない。各トークンが参照する専門家は一部でも、ルーターが選択し得る全専門家の重みを、GPU、CPU、分散ストレージのいずれかに保持する必要がある。 10



重みだけを単純換算した理論メモリー量は次のようになる。これは量子化を公式にサポートしているという意味ではなく、2.4兆という公開パラメータ数からの算術上の下限である。KVキャッシュ、ルーター、埋め込み、視覚モジュール、通信バッファ、アクティベーション、量子化メタデータ、冗長化を含まない。 11

仮定精度	重みのみの容量	80GBアクセラレーターの理論最小数	実務的含意
FP16／BF16相当	約4.8TB	60基	実際には60基を大きく上回り得る
FP8／INT8相当	約2.4TB	30基	公式FP8成果物は未公開
INT4相当	約1.2TB	15基	公式4bit量子化・品質評価は未公開

95Bの有効パラメータだけをFP16相当で読む場合のデータ量は約190GB／トークン相当になるが、実際の通信量やキャッシュ効率はテンソル並列・専門家並列・ルーティング方式に依存する。レイヤー数、専門家数、Top-k、共有専門家の有無、通信トポロジーが未公表であるため、H100、H200、B200、Ascendなどでのtokens/sや最適GPU構成を推定する根拠はない。

Model Studioの東京リージョンでは、画像・テキスト・動画入力、Function Calling、構造化出力、コンテキストキャッシュに対応する。一方、東京では組み込みWeb検索、バッチ推論、モデル調整は提供されない。入力上限は991,808トークン、思考モードでは983,616トークンである。リージョン上限は30,000 RPM、5,000,000 TPMとされるが、これは個別リクエストの速度やレイテンシ保証ではない。 12

性能評価とモデル比較

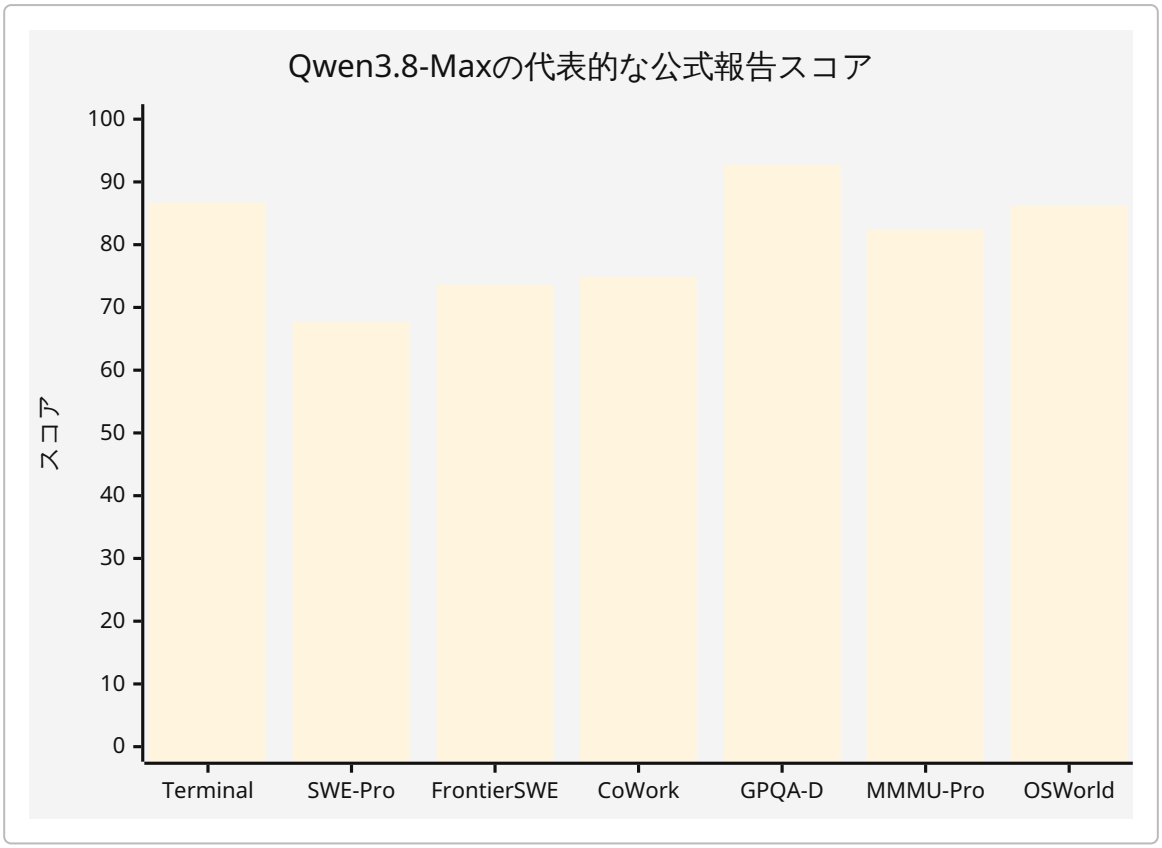
Alibabaの公式比較表は、2026年時点のOpus 4.8、Fable 5、GPT-5.6 Sol、Qwen3.7-Maxなどを主な比較対象としている。以下は公式ブログに埋め込まれた表から代表的な指標を転記したものであり、各モデルで評価ハーネス、タイムアウト、最大出力長、利用可能なツールが完全には統一されていない。特にClaude系には

Claude Code、GPT系にはCodex、QwenにはQwen用ハーネスが使われる場合があるため、モデル本体だけの能力比較ではなく、**モデルとエージェント実行環境の合成評価**として読むべきである。 2

ベンチマーク	Opus 4.8	Fable 5	GPT-5.6 Sol max	Qwen3.7-Max	Qwen3.8-Max
Terminal-Bench 2.1	84.6	84.6	**88.8**	74.5	86.6
SWE-bench Pro	69.2	**80.0**	64.6	60.6	67.7
FrontierSWE	70.0	**88.8**	未掲載	40.7	73.5
PaperBench	80.3	88.8	90.5	64.8	**93.0**
CoWorkBench	72.3	**75.9**	71.5	64.6	74.8
Toolathlon-Verified	76.2	**77.9**	74.9	49.7	72.5
GPQA Diamond	92.0	92.6	**94.1**	92.4	92.6
IFBench	62.2	63.5	72.7	79.1	**82.8**
LongBench v2	**69.1**	未掲載	67.1	65.3	66.3

この表からは、旧Qwen3.7-Maxに対する改善は明確である。PaperBenchは64.8から93.0、FrontierSWEは40.7から73.5、Toolathlon-Verifiedは49.7から72.5となっている。一方で、SWE-bench ProはFable 5より12.3ポイント低く、Toolathlonも5.4ポイント低い。したがって「コーディング全般で最強」ではなく、研究再現、指示追従、フロントエンド、長期的ワークフローに強い一方、特定の実リポジトリ修正や厳格なツール課題では他モデルに及ばない、というのが表から導ける妥当な評価である。 2

マルチモーダル評価では、Qwen3.8-MaxはMMMU-Pro 82.3、MathVision 95.2／Pythonツール利用時97.7、OSWorld-Verified 86.1、AndroidWorld 85.3、OmniDocBench 1.5で92.1、VideoMMEで90.4などを報告している。MMMU-ProはGPT-5.6 Solの83.0にわずかに届かないが、Fable 5の81.2を上回る。OSWorld-Verifiedは公式比較表の対象中で最高だった。ただし、これらもAlibaba側の評価であり、画像前処理、動画フレーム抽出、プロンプト、ツール構成が独立に再現されていない。 2



このグラフの各指標は異なる課題・採点方式を用いるため、平均して単一の総合点を作ることはできない。特にGPQAとOSWorld、SWE-benchを同じ尺度の能力として解釈してはならない。

人間選好評価では、LMArenaの結果が最も重要である。

LMArena部門	順位	スコア	投票数	状況
Text Arena	5位	1496±10	3,327	Preliminary
Vision Arena	2位	1305±9	5,038	Preliminary
WebDev Arena	4位	1668±18	1,563	Preliminary

Text Arenaでは4位のClaude Opus 4.6が1497±4、Qwen3.8-Maxが1496±10であり、統計的には明確な差があるとは言いにくい。Visionでは1位Fable 5が1318±9、Qwenが1305±9である。WebDevではQwenの1668±18は強いが、投票数が少なく信頼区間も広い。これらは「フロンティアモデル群に入った」ことを支持する一方、「安定して世界5位」「視覚で確定2位」と断定するには早い。 3

ユーザー指定の旧世代・他系列モデルとの比較可能性は次のとおりである。

比較対象	Qwen3.8-Maxとの厳密な比較状況	結論
Qwen3.8 非Max	独立した非Maxモデルのモデルカード・同一評価値は未公開	比較不能
Qwen2	公式発表の同一ハーネス表に含まれない	世代差は大きいと考えられるが数値比較は不適切

比較対象	Qwen3.8-Maxとの厳密な比較状況	結論
GPT-4o	同一時点・同一ハーネスの公式値なし	歴史的スコアとの直結は不適切
GPT-4o-mini	同上	価格帯・用途が異なり直接比較困難
Llama 3 70B	同上	オープンウェイト性と必要ハードウェアは比較可能だが性能表は非同条件
Claude 3	同上	旧世代スコアとの直接比較は非推奨
Claude 4	最新のOpus 4.8等とは公式表・LMArenaで比較あり	Qwenは一部課題で勝ち、一部で負ける

MMLU、HELM、LLM-Score、HumanEvalについては、Qwen3.8-Maxの正式な公開値を確認できない。安全性・毒性についてもRealToxicityPrompts、ToxiGen、HarmBench、JailbreakBenchなどの結果は公表されていない。堅牢性については、LongBench v2やMRCR v2など長文評価はあるが、敵対的プロンプト、分布外入力、文字化け、言語混在、ツール出力汚染、プロンプトインジェクションへの体系的評価ではない。 2

独立レビュー、市場、エコシステム

独立検証は発表直後で極めて限定的であり、プレビュー版と正式版を区別する必要がある。

Trilogy AIの単一タスクによるブラインド評価では、Qwenは22回のゲートウェイ要求と44回のツール呼び出しを行い、ツール失敗はゼロ、ツール利用評価は10点中9点だった。一方、354件のリポジトリ引用のうち2件は行参照が誤り、7つの主張群はコードから証明できる範囲を超えていた。これは、検索・ツール実行・引用量は優秀でも、**引用が存在することと、その引用が主張を完全に立証することは別**であることを示す有用な事例である。 13

個人レビューでは、システム設計の過剰複雑化を避ける能力、バックエンド・セキュリティ設計、フロントエンド、動画制作を高く評価する報告がある。一方、速度評価は一致していない。あるAPI利用者は「モデル規模の割に速い」と述べる一方、Web／サブスクリプション利用者は遅さを報告している。別のレビューでは、単一のWeb制作課題に30分以上かかった。この相違は、リージョン、混雑、エージェントの反復回数、`reasoning_effort`、ツール実行時間、APIとWeb UIの違いに強く依存する可能性を示すが、公式のTTFT、tokens/s、p50／p95値がないため原因を特定できない。 14

プレビュー版では、長時間のツール利用に関する具体的な不具合報告がある。Anthropic互換APIを約550Kトークンまで使用したセッションで、ツール名とJSON引数の先頭が失われたという報告があり、別の200ターン超・約187K〜218Kトークンのセッションでは、構造化ツール呼び出しがXML形式の通常テキストとして出力され、ツールが実行されなかった。いずれもGitHub上のユーザー報告であり、正式版に残っていることは確認されていないが、長期エージェントの本番導入前に再現試験すべき高優先度のシナリオである。 5

市場面では、Qwen3.8-Maxは同等クラスのフロンティアモデルに対して価格攻勢をかけている。QwenCloudの表示価格は入力100万トークン2ドル、出力6ドルである。LMArena上のClaude Opus 4.6は5ドル／25ドル、Fable 5は10ドル／50ドルであり、少なくとも公称API単価ではQwenが大幅に低い。性能差が小さい業務では、モデルルーティングや第2モデルとして採用される可能性が高い。 15

Alibaba Cloudでは北京、シンガポール、東京、フランクフルト、米国バージニアの各リージョンで提供され、OpenAI互換・Anthropic互換APIが用意される。QwenCloudのほか、OpenRouterでも100万トークン、2ドル／6ドルのモデルとして掲載されている。Qwen Code、Claude Code、Codex、Qoder、OpenClawなど既存のエージェント環境に接続できる点は、移行コストを下げる。もっとも、API形式の互換性は、ツール

呼び出し、思考内容、ストリーミングイベント、エラーコード、JSON制約まで完全に同一であることを保証しない。[16]

オープンウェイト化の影響は二面性がある。一方では、Maxクラスの重みが公開されれば、研究機関や大企業が蒸留、量子化、特定言語・業務への適応を研究できる。他方、4bitでも重みだけで約1.2TBとなるため、単一ワークステーションや一般的な8GPUサーバーでの完全常駐は困難である。Qwen3.8-Maxの公開は「誰でもローカル実行できる民主化」というより、**大規模クラスタを持つクラウド事業者、国家研究機関、大企業に対するフロンティア級ベースモデルの供給**という意味が強い。

Qwen3.8-Max固有の企業採用事例は、正式発表翌日の段階では確認できない。Alibaba Cloudが掲載する既存の顧客事例は、Qwen、Wan、Model Studio全般の導入事例であり、Qwen3.8-Maxの本番採用実績として数えるべきではない。Reutersは発表時にAlibaba株の上昇を報じているが、これは市場の短期反応であり、売上、継続利用率、企業契約数の証拠ではない。[17]

リスクとガバナンス

最も重大なリスクは、安全性情報が能力情報に比べて著しく不足していることである。公開資料には、毒性、偏見、虚偽情報、児童安全、性的コンテンツ、テロ、化学・生物、サイバー攻撃支援、モデル抽出、学習データ漏えい、プロンプトインジェクション、脱獄に関する体系的な評価がない。モデルが長期間にわたりコード、ブラウザー、OS、外部ツールを操作できるとするなら、通常のチャットモデルよりも、誤操作や悪意ある入力が現実の副作用につながりやすい。[18]

特にFunction Callingを利用する場合、次の統制が必要になる。

リスク	典型的な失敗	必須の対策
過剰な権限	本番DB、クラウド、Gitへの破壊的変更	最小権限、読み取り専用既定、短命資格情報
プロンプトインジェクション	Web、PDF、Issue内の命令を優先	外部コンテンツを非信頼データとして分離
ツール形式崩壊	JSONではなくXML・平文を出力	スキーマ検証、拒否、再試行、フォールバック
長期ドリフト	数百ターン後に方針や制約を忘れる	状態の外部管理、定期的な再要約と再認証
コスト暴走	100万トークン文脈を反復送信	トークン予算、キャッシュ、時間上限
データ漏えい	機密文書を外部APIへ送信	DLP、匿名化、リージョンと契約条件の確認
誤った引用	参照はあるが主張を裏付けない	引用範囲検証、二次判定モデル、人間レビュー

中国国内で一般向け生成AIサービスを提供する場合、中国の「生成式人工智能服务管理暂行办法」が適用され得る。同規則は中国国内の公衆向け生成コンテンツサービスを対象とし、合法性、個人情報、知的財産、コンテンツ管理などを求めている。AlibabaのAPIを利用すること自体と、利用企業が中国国内で一般向けサービスを提供することは別の法的問題であり、提供地域、利用者、データフロー、サービス形態ごとに確認が必要である。[19]

EUではGPAIモデルに関する透明性、著作権、技術文書等の義務が2025年8月から適用され、欧州委員会の執行権限は2026年8月2日から本格化した。システミックリスクに該当するGPAIには、標準化されたモデル評価、敵対的テスト、リスク軽減、重大インシデント報告、サイバーセキュリティが求められる。Qwen3.8-MaxをEUで提供・統合する場合、Alibaba側のモデル提供者義務だけでなく、下流システムの提供者・導入者としての文書化、透明性、人間監督、用途別義務を検討する必要がある。 20

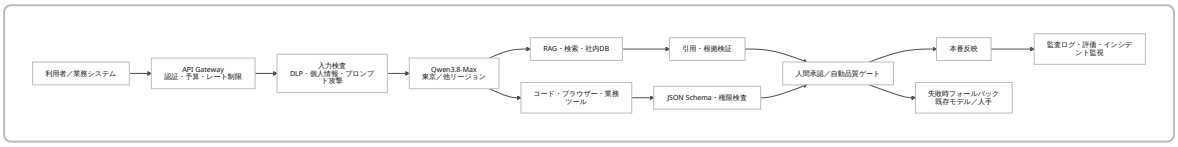
日本では、経済産業省・関係機関が2026年3月に「AI事業者ガイドライン第1.2版」を取りまとめている。Qwen3.8-Maxを導入する企業は、法令遵守だけでなく、リスクベースの評価、説明可能性、セキュリティ、人間中心、継続監視、インシデント対応を自社のAIガバナンスに組み込むべきである。 21

供給網については、2.4Tモデルの自社運用に大量の先端アクセラレーターと高速相互接続が必要になる一方、米国の先端計算品目に対する輸出管理は継続的に変化している。BISは2026年5月のガイダンスで、D:5国・マカオに本社または最終親会社を持つ事業体向けの一定の先端計算品目について、事業体が第三国に所在していてもライセンス要件が継続すると明確化した。これはQwen3.8-Maxが特定ハードウェアで訓練されたことを示すものではないが、中国系モデルの訓練・ホスティング能力、GPU調達価格、クラウド提供地域に影響し得る構造的リスクである。 22

実務導入ガイドとコスト

推奨度が高い用途は、既存コードベースの調査・設計レビュー、フロントエンド試作、テスト生成、長文書・PDF・動画の分析、調査エージェント、定型オフィス業務、GUI操作の検証、低コストな第2モデル／フォールバックモデルである。特に、成果物をテスト、静的解析、スキーマ検証、引用照合で機械的に確認できる業務と相性がよい。公式・LMArenaの結果は、コーディング、視覚、OS操作、指示追従での強さを支持している。 23

現段階で単独利用を避けるべき用途は、医療診断、信用・保険判断、法的最終判断、自動採用・人事評価、工場・電力・交通などの安全制御、無監督の本番インフラ変更、攻撃的サイバー操作、大額決済である。これは当該能力が低いからではなく、モデルカード、安全評価、正式な堅牢性評価、長時間ツール利用の信頼性データが不足しているためである。



API単価は、QwenCloudで入力2ドル／100万トークン、出力6ドル／100万トークンと表示される。Alibaba Cloudの東京リージョン価格は入力12人民元／100万、出力36人民元／100万、キャッシュヒット入力1.5人民元／100万である。価格はキャンペーン、契約、リージョンで変わり得る。 24

課金項目	1,000トークン当たり	100万トークン当たり
QwenCloud入力	0.002ドル	2ドル
QwenCloud出力	0.006ドル	6ドル
東京リージョン入力	0.012人民元	12人民元
東京リージョン出力	0.036人民元	36人民元
東京リージョン・キャッシュヒット入力	0.0015人民元	1.5人民元

概算例は次のとおりである。

1回の処理量	QwenCloud概算	東京リージョン概算
入力10万＋出力2万トークン	0.32ドル	1.92人民元
入力100万＋出力20万トークン	3.20ドル	19.20人民元
入力1万＋出力2千トークン	0.032ドル	0.192人民元

画像や動画が何トークンに換算されるか、推論チェーンが請求対象出力にどう反映されるか、ツール反復が何回発生するかによって実額は大きく変わる。100万トークンを毎ターン再送するとコストと遅延が急増するため、固定システムプロンプト、共通資料、コードベース索引にはコンテキストキャッシュを利用し、会話履歴は外部状態と要約に分離するのが望ましい。[25]

ファインチューニングについては、Model StudioがQwen3.8-Maxの「モデル調整非対応」を明示しているため、現時点のAPIファインチューニング費用は**未提供・未価格設定**である。重み公開後のLoRA／QLoRA費用も、モデル構成、対応精度、専門家並列方式、利用可能な量子化、ライセンスが不明なため、責任ある数値見積もりはできない。4bit重みだけで理論上約1.2TBであることを踏まえると、通常の単一8GPUノードでのLoRAは困難であり、専門家オフロードまたは複数ノードが必要になる可能性が高い。[26]

既存モデルからの移行は、単なるモデル名の置換ではなく、段階的なカナリア方式で行うべきである。

移行項目	推奨措置
GPT-4o／OpenAI APIから	OpenAI互換エンドポイントを利用しつつ、ストリーミング、JSON Schema、ツール呼び出し、エラー処理を再試験
Claudeから	Anthropic互換APIを利用できるが、長時間ツールセッションで構造崩壊試験を実施
Qwen3.7から	同じ評価セットで3.7／3.8を並列実行し、品質向上分とトークン増加分を比較
長文RAGから	最初から100万トークンを投入せず、検索・再ランキング・引用確認と組み合わせる
コーディングエージェント	読み取り専用から開始し、テスト成功時のみ書き込み、デプロイは人間承認
日本語業務	敬語、固有名詞、契約文、表、文字コード、日英混在を社内セットで評価
推論設定	通常処理は`low`または`medium`、難問のみ`xhigh`をルーティング
障害対策	タイムアウト、最大反復数、最大費用、既存モデルへのフォールバックを設定

導入判断としては、短期的にはAPIで評価し、少なくとも数週間の本番類似テストを行うのが妥当である。重み、ライセンス、モデルカード、量子化成果物、推論エンジン対応が公開されるまでは、自社ホスティングの設備購入や大規模移行を確定すべきではない。

優先ソースと原典

優先度	原典	公開・更新日	URL
最優先	Qwen公式「Qwen3.8-Max: A New Bar for Coding and Cowork」	2026-08-02表示／正式発表は8月3日	https://qwen.ai/blog?id=qwen3.8
最優先	Alibaba Cloud正式発表	2026-08-03	https://www.alibabacloud.com/blog/alibaba-unveils-qwen3-8-max-its-largest-and-most-capable-flagship-model-to-date_603420
最優先	Alibaba Cloud版技術・APIブログ	2026-08-03	https://www.alibabacloud.com/blog/qwen3-8-max-a-new-bar-for-coding-and-cowork_603421
最優先	Alibaba Cloud Model Studio Qwen3.8-Max仕様・価格	2026-08-04確認	https://help.aliyun.com/zh/model-studio/qwen3-8-max
高	Model Studioモデル一覧・リージョン仕様	2026-08-04確認	https://help.aliyun.com/zh/model-studio/models
高	QwenCloudモデル・デモ・API入口	2026-08-04確認	https://www.qwencloud.com/
高	Qwen Code設定文書	2026-08-04確認	https://help.aliyun.com/zh/model-studio/qwen-code
高	Qwen Hugging Face組織ページ	2026-08-04確認、Qwen3.8-Max重み未掲載	https://huggingface.co/Qwen
高	Qwen公式GitHub	2026-08-04確認、Qwen3.8-Max固有実装未掲載	https://github.com/QwenLM
高	LMArena Text Leaderboard	2026-08-04確認	https://arena.ai/leaderboard/text
高	LMArena Vision Leaderboard	2026-08-04確認	https://arena.ai/leaderboard/vision
高	LMArena WebDev Leaderboard	2026-08-04確認	https://arena.ai/leaderboard/code
中	Reuters発表報道	2026-08-03	https://www.reuters.com/business/retail-consumer/alibaba-unveils-its-most-capable-ai-model-date-not-far-behind-moonshots-size-2026-08-03/
中	Trilogy AI独立タスク評価	2026-08確認	https://trilogyai.substack.com/p/qwen-38-max-benchmark-how-it-compares

優先度	原典	公開・更新日	URL
中	Qwen3 GitHubツール呼び出し不具合報告	2026-08-01	https://github.com/QwenLM/Qwen3/issues/1886
中	Qwen Code長文セッション不具合報告	2026-07-29	https://github.com/QwenLM/qwen-code/issues/8003
規制	中国「生成式人工智能服务管理暫行办法」	2023-07-13	https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm
規制	EU AI Act公式概要	2026-08-04確認	https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
規制	日本「AI事業者ガイドライン第1.2版」	2026-03-31	https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20260331_report.html
供給網	米国BIS先端計算品目ガイダンス	2026-05-31	https://www.bis.gov/media/documents/bis-guidance-may-31-2026.pdf

現時点で未指定の中心項目は、学習データの規模・構成、トークナイザー、内部レイヤー構成、専門家数とルーティング、訓練ハードウェア、訓練計算量、FP16／FP8／4bit成果物、レイテンシ、LoRA手順、ライセンス、安全性評価である。これらについては、翌週予定のモデル重み・モデルカード公開後に初めて、再現可能性、自社運用可能性、法務・安全性を含む最終評価が可能になる。

---

1 6 9 10 11 18 23 Alibaba Unveils Qwen3.8-Max: Its Largest and Most Capable Flagship Model to Date - Alibaba Cloud Community
https://www.alibabacloud.com/blog/alibaba-unveils-qwen3-8-max-its-largest-and-most-capable-flagship-model-to-date_603420

2 4 7 16 25 Qwen3.8-Max: A New Bar for Coding and Cowork - Alibaba Cloud Community
https://www.alibabacloud.com/blog/qwen3-8-max-a-new-bar-for-coding-and-cowork_603421

3 LLM Leaderboard - Best Text & Chat AI Models Compared
https://lmarena.ai/leaderboard/text

5 Tool-call emission loses its prefix — only the tail of the argument JSON arrives (qwen3.8-max-preview)・Issue #1886・QwenLM/Qwen3・GitHub
https://github.com/QwenLM/Qwen3/issues/1886

8 [2505.09388] Qwen3 Technical Report
https://arxiv.org/abs/2505.09388?utm_source=chatgpt.com

12 26 qwen3.8-max 模型信息-大模型服务平台百炼(Model Studio)-阿里云帮助中心
https://help.aliyun.com/zh/model-studio/qwen3-8-max

13 Qwen 3.8 Max Benchmark: How It Compares With Kimi K3
https://trilogyai.substack.com/p/qwen-38-max-benchmark-how-it-compares

14 My thoughts on Qwen 3.8 Max Preview : r/Qwen_AI
https://www.reddit.com/r/Qwen_AI/comments/1v6bnu9/my_thoughts_on_qwen_38_max_preview/

15 24 QwenCloud - AI-Native Models, Tools & Apps Platform - Ready Out of the Box
https://www.qwencloud.com/

17 Alibaba unveils its largest AI model yet, DeepSeek's latest model is ultra-low cost | Reuters
https://www.reuters.com/business/retail-consumer/alibaba-unveils-its-most-capable-ai-model-date-not-far-behind-moonshots-size-2026-08-03/

19 生成式人工智能服务管理暂行办法
https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm?utm_source=chatgpt.com

20 AI Act | Shaping Europe’s digital future
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

21 AI事業者ガイドライン
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20260331_report.html?utm_source=chatgpt.com

22 bis.gov
https://www.bis.gov/media/documents/bis-guidance-may-31-2026.pdf