Qwen4-Max-Preview 徹底調査レポート：仕様・アーキテクチャ・経済性・実測評価の全貌

Gemini 4 Carbon

Qwen4-Max-Preview（モデルID：qwen4-max-preview、課金エイリアス：qwen4.0-max-preview）は、Alibaba Cloud（Qwenチーム）が2026年11月11日10時00分（中国標準時 UTC+8／日本時間11時00分）に公開した、次世代「Qwen4」シリーズの最上位マルチモーダルフラッグシップモデルの先行試用版である[1]。最大100万トークンの入力コンテキスト長と、前世代Qwen3.8-Max（131,072トークン）の3倍となる最大393,216トークンの出力上限（内部思考プロセスと最終回答の共有枠）を備え、テキスト・画像・動画の統合理解、既定で最高強度（xhigh）に設定されたハイブリッド推論、Function Callingや組み込みWeb検索・コードインタープリタ、厳格なJSON Schema構造化出力、ならびにOpenAIおよびAnthropic互換APIプロトコルをサポートする[1]。複雑なソフトウェアエンジニアリング、フロントエンド構築、専門的なオフィス業務、長期的な自律エージェント処理を主眼に設計されており、chat.qwen.ai（Qwen Studio／Qwen Chat）や公式アプリから無料で試用できるほか、開発者向けにはサブスクリプション型サービス「Token Plan」を通じてClaude CodeやCursorなどの対話型コーディングツールへ接続可能である[6]。

ただし、2026年11月15日時点で評価・導入を検討するにあたり、重要な制約と情報非開示の前提が存在する。第一に、本モデルの提供は**2026年11月25日24時00分（UTC+8**／日本時間**11月26日午前1時00分）**までの約**15日間（350時間）**限定であり、公式ドキュメントによれば期間中はモデルの重みが固定（変更なし）された後、期限到来とともにプレビューAPI自体が退役（提供終了）する[2]。第二に、Alibabaは今回のリリースにおいてパラメータ数、アーキテクチャ詳細、モデルカード、公式ベンチマークスコア（**GPQA**、**SWE-bench**、**AIME**等）、オープンウェイト（重みデータ）、および従量課金**API**のトークン単価を一切公表していない[5]。第三に、公式APIアクセスは月額クレジット制の「Token Plan（個人版・チーム版）」に限定されており、その利用規約上、人間が操作するコーディング・エージェントツール内での対話的利用のみが許可され、自社アプリケーションのバックエンド組み込み、自動化スクリプト、バッチ処理への利用は禁止（違反時はAPIキー一停止・契約凍結の対象）されている[6]。

公式ベンチマークが不在の中、公開直後から技術メディアや開発者コミュニティが実施した複数の第三者検証および実務ストレステストを総合すると、Qwen4-Max-Previewは「高い基礎精度とハルシネーション耐性を備えた重量級推論モデル」と評価されている[12]。財務諸表の横断検証や約500KBのログからの長文脈検索、自律的なテスト駆動型コードリファクタリング、微細な視覚特徴の検知と物理・法務知識の統合において堅実な結果を示し、コミュニティによる小規模検証では思考強度xhigh設定時にAnthropicのClaude Opus 5.5（Medium設定）に近い品質、また汚染排除型のサンドボックスエージェント評価ではDeepSeek-V4.1-Flashと総合解決数で拮抗しつつ帰納的な規則発見で優位に立つ結果が報告されている[12]。その反面、深い推論に伴う実行時間の長さ（複雑なタスクで8分～41分超）や、max_tokensパラメータの仕様変更による出力切断リスク、サーバー側システムルールのコンテキスト漏洩バグなど、開発者が事前に把握しておくべき特有の挙動も確認されている[5]。

1. リリース概要と提供ライフサイクル：15日間の公開ストレステ

スト

AlibabaのQwen公式Xアカウント（@Alibaba_Qwen）は、2026年11月6日に「The becoming begins」というティザー投稿を行った後、5日後の11月11日10時01分（北京時間）に「Qwen4 is coming」と宣言し、シリーズ第1弾としてQwen4-Max-Previewを同時公開した[8]。公開前の時点では、AlibabaのAIコーディング製品「Qoder」上で「Qoduck」と呼ばれる正体非公開のステルスモデルが日次無料枠付きで提供されており、Qwen4との関連は推測にとどまっていたが[21]、海外技術ニュースの報道によれば11月11日の正式発表に合わせてQoder側が「Qoduck」の正体がQwen4-Max-Previewであったことを確認し、同時にQoderの価格を80%引き下げたと伝えられている[22]。

今回のリリース形態で特筆すべきは、Alibabaが従来のAIモデル発表における慣行であった「テクニカルレポートと公式ベンチマーク比較表の同時公開」を見送った点である[12]。公式アナウンスは開発者に対して「現実の難題をQwen4に投げて試してほしい」と呼びかけるにとどまり、独立系評価機関であるArtificial AnalysisやBenchLM、AI Release Trackerなどの主要リーダーボードにも、ローンチ時点でQwen4-Max-Previewの計測スコアは掲載されていない（なお同時期のArtificial AnalysisにおけるClimbベンチマーク評価ではClaude Fable 5.2がWSR 57%、Gemini 4 Carbonが54%、Claude Opus 5.5が53%を記録しているが、Qwen4-Max-Previewは対象外となっている）[5]。

この「ベンチマーク非公表・15日間限定・重み固定プレビュー」という提供方式は、前世代のQwen3.8でも採用された段階的ロールアウト戦略を踏襲・発展させたものとみられる[12]。中国メディアSohu Techの分析や開発者コミュニティの指摘によれば、静的な公開ベンチマークの順位競争から軸足を移し、Token PlanやQoder、QwenWork（千问办公）などの実務環境を通じて実際のコードベースや業務課題におけるワークフロートレースを収集し、正式版（General Availability：GA）公開前の事後学習（Post-Training）に反映させる狙いがうかがえる[12]。前世代Qwen3.8-Maxにおいても2週間のプレビュー版から正式版チェックポイントへの移行時に顕著な性能改善が報告されており[18]、11月25日24時（UTC+8）に退役するqwen4-max-previewは最終製品ではなく「次世代フラッグシップの公開ストレステスト版」として位置づけるのが妥当である[12]。

また、「Qwen4」ファミリー全体の構成については、カンファレンス資料に関する報道を通じて最上位のQwen4-Max、バランス型のQwen4-Plus、効率重視のQwen4-Flash、およびオープンウェイト枠と想定されるQwen4-27Bの4名称が流通している[21]。しかし、2026年11月15日時点で実際に呼び出し可能なのはQwen4-Max-Previewのみであり、正式版Qwen4-Maxを含む各ティアの公開日、ライセンス体系（Qwen3.8世代では27BがApache 2.0、Flash-Nextが独自コミュニティライセンス、Maxがカスタムライセンスと階層別に異なっていた）、およびオープンウェイト提供の有無は公式には未発表である[14]。なお、2026年9月のAlibaba Cloud発表で言及された「5兆〜10兆パラメータ」という見通しは将来の「Qwen4.5」および「Qwen5」シリーズに関するものであり、Qwen4自体の規模を指す数値ではない[21]。

2. 技術仕様と前世代フラッグシップとの比較

Qwen4-Max-Previewの確定技術仕様における最大の変更点は、100万トークンの入力コンテキスト長を維持しつつ、1回のリクエストあたりの最大出力上限を前世代Qwen3.8-Max（131,072トークン）の3倍にあたる**393,216**トークンへと拡張した点にある[5]。

Qwen4-Max-Previewは最大出力枠を前世代比3倍の39.3万トークンへ拡張し思考と回答で共有する



上段は公式ドキュメントで確認されたQwen4-Max-Previewの基本仕様を示す。下段はQwen世代間の最大出力トークン上限を比較したもの。Qwen4-Max-Previewでは最大393,216トークンの枠を内部思考（既定値xhigh）と最終回答が動的に共有する設計となっており、従来の最終回答のみを想定した小さな上限設定では出力が途中で切断されるリスクがある。

データ出典: Alibaba Cloud Model Studio, QwenCloud Documentation, Local AI Zone, eesel AI, Yotta Labs

この393,216トークンという出力枠の拡大は、デフォルトでxhigh（最高強度）に設定された「深い思考（Deep Thinking）」と長大なコード・文書生成を単一パスで両立させるための設計である[5]。複数ファイルにまたがる大規模リポジトリの改修や、3Dライブラリ・アセットを単一HTMLファイル内に埋め込む数千行規模のフロントエンド実装では、モデルが実装前の設計やエッジケース検証に多量の思考トークン（reasoning_content）を費やしたうえで、完全な実行可能コードを出力する必要がある[1]。従来の131,072トークン上限と比較して3倍の共有バジェットが確保されたことで、長期的な計画維持と一括出力の余地が大幅に広がっている[1]。

マルチモーダル理解に関しては、テキストに加えて画像と動画の入力を受け付け、UIスクリーンショット解析、複数ページ文書のOCR・構造化抽出、オブジェクト位置特定、および動画内のタイムスタンプ付きイベント分析に対応する[6]。公式ドキュメントが参照先として指定する視覚理解ガイドライン（視覚モデル共通仕様およびQwen3.8系基準）では、最大64本までの動画入力（フレーム抽出レートfpsは既定2.0、0.1〜10の範囲で調整可；Qwen3.8系では公開URL経由で最大2GB・長さ2秒〜2時間）や8K以内推奨の画像入力（Qwen3.8系では1トークン＝32×32ピクセル換算）が規定されている

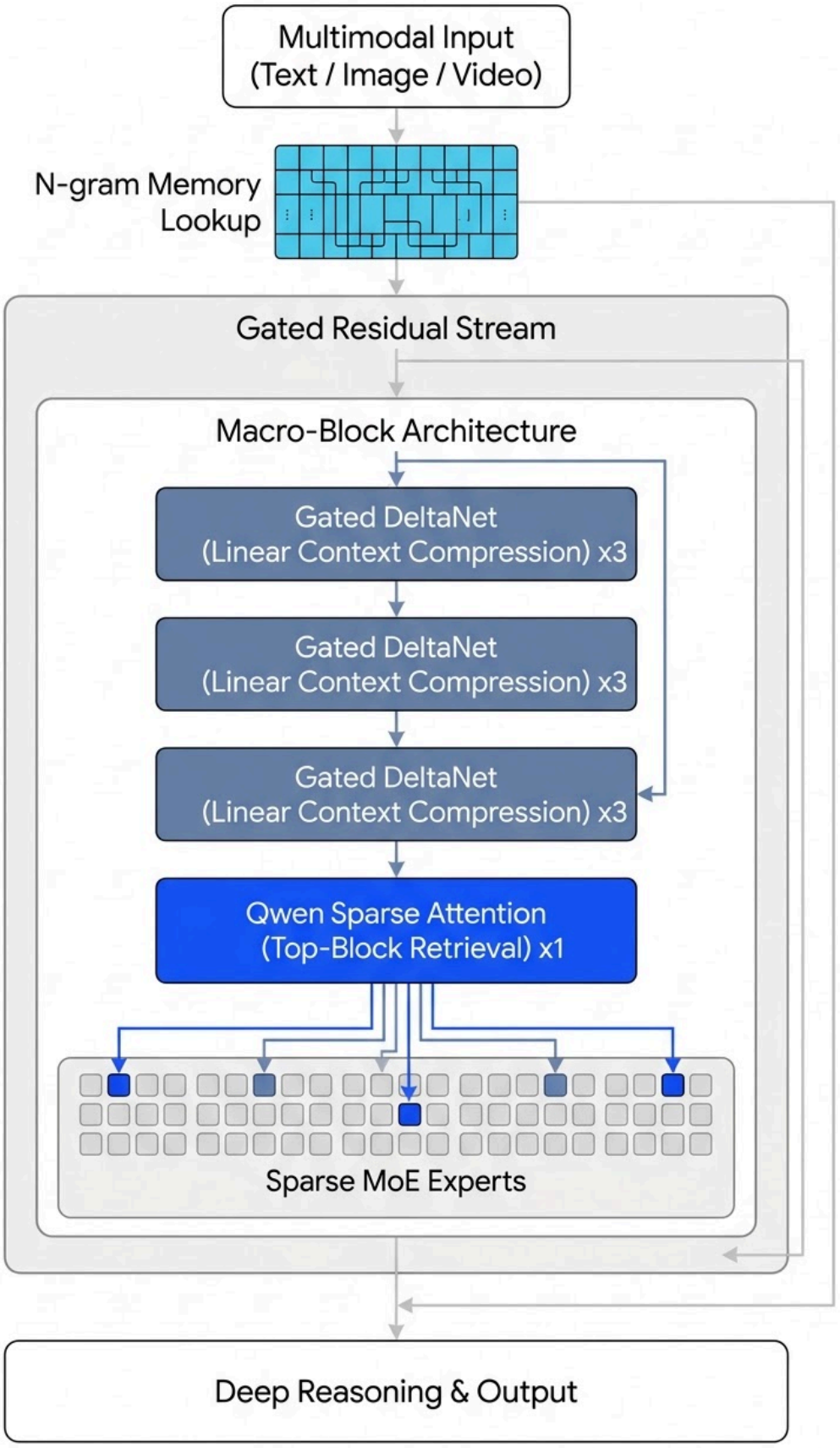
一方、視覚モデル共通の制約として動画ファイル内の音声トラック理解には対応しておらず、TXT・Word（.docx）・PDFなどのテキスト系ファイルも直接視覚エンドポイントには渡せないため、ページ単位の画像へ変換するか文書用のQwen-Longを併用する運用が案内されている[6]。
外部ツール連携および出力制御では、標準的なFunction Calling（独立タスクの並列呼び出しparallel_tool_calls: true対応）に加え、OpenAI Responses API経由で自動発動する5種の組み込みモデルツール（Web検索web_search、テキストからの画像検索web_search_image、逆画像検索image_search、Webスクレイピングweb_extractor、コード実行環境code_interpreter）を利用できる[6]。構造化出力は柔軟なJSON Objectモードと、PydanticやZodと連携して型・必須キーを強制するJSON Schema（strict: true）モードの双方をサポートするが、画像・動画などのマルチモーダル入力を含むリクエストではJSON Schemaがサポートされず、自動的にJSON Objectへフォールバックする仕様となっている[6]。
現行の正式フラッグシップであるQwen3.8-Maxおよびアーキテクチャ先行公開版Qwen3.8-Flash-Nextとの仕様比較は以下の通りである。

比較項目	**Qwen4-Max-Preview**	**Qwen3.8-Max（0902含む）**	**Qwen3.8-Flash-Next**
モデル**ID**	qwen4-max-preview（別名qwen4.0-max-preview）[4]	qwen3.8-max / qwen3.8-max-0902[7]	オープンウェイト / qwen4_exp系[30]
位置づけ・状態	次世代最上位の15日間限定プレビュー（11/25退役）[6]	現行の正式フラッグシップ（2026年8月GA）[11]	Qwen4世代アーキテクチャの実験的先行公開版（8月公開）[14]
総パラメータ / アクティブ数	非公開[14]	総2.4兆（2.4T）MoE / アクティブ約950億（95B）[14]	保存時約180B（125B MoE + 51B N-gram + 4B MTP）/ アクティブ**6B**[30]
入力コンテキスト長	1,000,000トークン[6]	1,000,000トークン[27]	最大1,000,000トークン（標準26.2万、静的YaRN）[30]
最大出力トークン	**393,216**トークン（思考＋回答の共有）[5]	131,072トークン[5]	思考最大262,144 ＋ 回答最大131,072[30]
入力モダリティ	テキスト・画像・動画[6]	テキスト・画像・動画[11]	テキスト・画像・動画（ネイティブマルチ

			モーダル）[14]
思考モード仕様	Hybrid（既定ON：low / medium / xhigh［既定］）[6]	API版はHybrid（既定ON）、OSS版2.4TはThinking-only[6]	Hybrid（既定ON：xhigh既定、Instructモード切替可）[30]
オープンウェイト公開	なし（クラウドホスト限定）[5]	あり（2026年8月12〜13日公開、カスタムライセンス）[14]	あり（Qwen Community License 1.0）[34]
API提供・料金形態	**Token Plan**限定（従量課金なし・単価非公表）[5]	従量課金API（入力 2.00/出力 6.00 per 1M）＋Token Plan[15]	セルフホストおよび各種低価格API（3.8-Maxの約12分の1水準）[30]
公式ベンチマーク	未公表[14]	ベンダー公表済（SWE-bench Pro 67.7%、Terminal-Bench 2.1 86.6等）[37]	ベンダー公表済（SWE-bench Pro 62.5、GPQA Diamond 91.7、LCB v6 91.9）[30]
自社アプリのバックエンド利用	不可（規約により対話ツール利用のみ）[5]	可能（従量課金API経由）[11]	可能（ライセンス準拠のセルフホスト・API）[31]

3. アーキテクチャの系譜：Qwen3.8-Flash-Nextから読み解くQwen4の設計思想

AlibabaはQwen4-Max-Preview自体の内部構造やパラメータ数を非開示としているものの、2026年8月26〜28日に公開したオープンウェイトモデルQwen3.8-Flash-Next（内部アーキテクチャ識別子：arch: qwen4_exp）について、「完全なQwen4ロールアウトに先立ち、**Qwen4**の基盤となるアーキテクチャを実験的に先行公開したもの」と公式に位置づけている[14]。フラッグシップであるQwen4-Max-Previewがこの基本設計をどの規模へスケールアップまたは調整しているかは未発表であるものの、公式に記録されているQwen4世代の設計方針は「より希薄（リーン）なアクティブ計算量、大規模メモリ層の統合、ネイティブ・マルチモーダル、100万トークン文脈の高速処理」の4点にある[14]。


Multimodal Input
(Text / Image / Video)
N-gram Memory
Lookup
Gated Residual Stream
Macro-Block Architecture
Gated DeltaNet
(Linear Context Compression) x3
Gated DeltaNet
(Linear Context Compression) x3
Gated DeltaNet
(Linear Context Compression) x3
Qwen Sparse Attention
(Top-Block Retrieval) x1
Sparse MoE Experts
Deep Reasoning & Output

Qwen3.8-Flash-Nextのテクニカルレポートに基づく技術解説（Local AI Zone等の詳細分析）によれば、このQwen4基盤アーキテクチャは以下の4つの技術要素で構成されている[30]。

第一に、**Gated DeltaNet**（**GDN**）と**Qwen Sparse Attention**（**QSA**）の「**3対1**」インターリーブ構造である[30]。従来のフルアテンション機構では系列長に対して計算量が二乗で増大するため、100万トークンのコンテキスト処理ではプリフィル（初期読み込み）の遅延とKVキャッシュの肥大化が課題となる。qwen4_exp設計では全48層を4層単位の「マクロブロック」12個に編成し、各ブロックの最初の3層に線形アテンションであるGDNを配して文脈を学習済みゲート付きの固定サイズ再帰状態へと圧縮する[30]。そして4層ごとに1層挿入される新機構「QSA」が、履歴をマイクロブロック単位に集約し、軽量なMulti-Query Attention（MQA）インデクサ（クエリ4ヘッド・共有キー1ヘッド）で重要度を推定して上位ブロック（固定予算：1層あたり最大512ブロックまたは2,048トークン）のみを選択的にアテンション計算にかける[30]。全層を4分岐のゲート付き残差構造（4-branch Gated Residual）で包むことで、100万トークン文脈において従来比8.6倍のスループットと精密な文脈検索性能の両立を図っている[30]。

第二に、トークンあたりの演算FLOPsを消費しない「大規模N-gram埋め込みメモリテーブル」の導入である[14]。Qwen3.8-Flash-Nextでは第2層に2,000万件のバイグラム・トライグラム項目を格納した510億パラメータ（51B）のN-gramテーブルを配置している[30]。これは行列乗算ではなく直前のトークン列に基づく決定論的なテーブル参照（ルックアップ）として動作するため追加の演算FLOPsが発生せず、GPUメモリへの非同期プリフェッチおよびホストシステムRAMへのオフロード（NVIDIA GPU環境）によって推論速度を維持したまま知識容量を拡張できる[30]。

第三に、高希薄な**MoE**（**Mixture-of-Experts**）と**Multi-Token Prediction**（**MTP**）ヘッドの組み合わせである[30]。先行版では512個のエキスパートのうちトークンごとに活性化されるのはルーティング10個＋共有1個（バックボーン125B中わずか6B、アクティブ率約4.8%）に絞られており、4BパラメータのMTPヘッドが投機的デコーディングを支援する[30]。前世代のQwen3.8-Maxでも総2.4Tに対してアクティブ約95B（約4%）という希薄設計が採られていたが[32]、コミュニティの初期検証でもQwen4-Max-Previewでは高推論設定時のトークン効率がさらに改善したとの声が上がっている[25]。

第四に、**Muon**オプティマイザと学習スケジュールの刷新である[30]。2次元線形マップ（QSA、GDN、MoEエキスパート）の学習にMuonオプティマイザ、埋め込み層やルーター等にAdamWを併用し、バッチサイズのウォームアップを省いた新スケーリング則を採用したことで、先行版ではオプティマイザのステップ数を18.8%削減し、学習コストをQwen3.7-Plusの約9分の1に抑えつつ、SWE-bench Proで62.5（比較対象のClaude Opus 4.6 Maxは53.4）、GPQA Diamondで91.7、LiveCodeBench v6で91.9というスコアを報告していた[30]。なお、一部のサードパーティブログではQwen4シリーズの学習に再帰的自己改善（Recursive Self-Improvement）が用いられたとの言及も見られるが、Alibaba公式からQwen4-Max-Previewの学習データや訓練手法に関する開示は行われていない[13]。

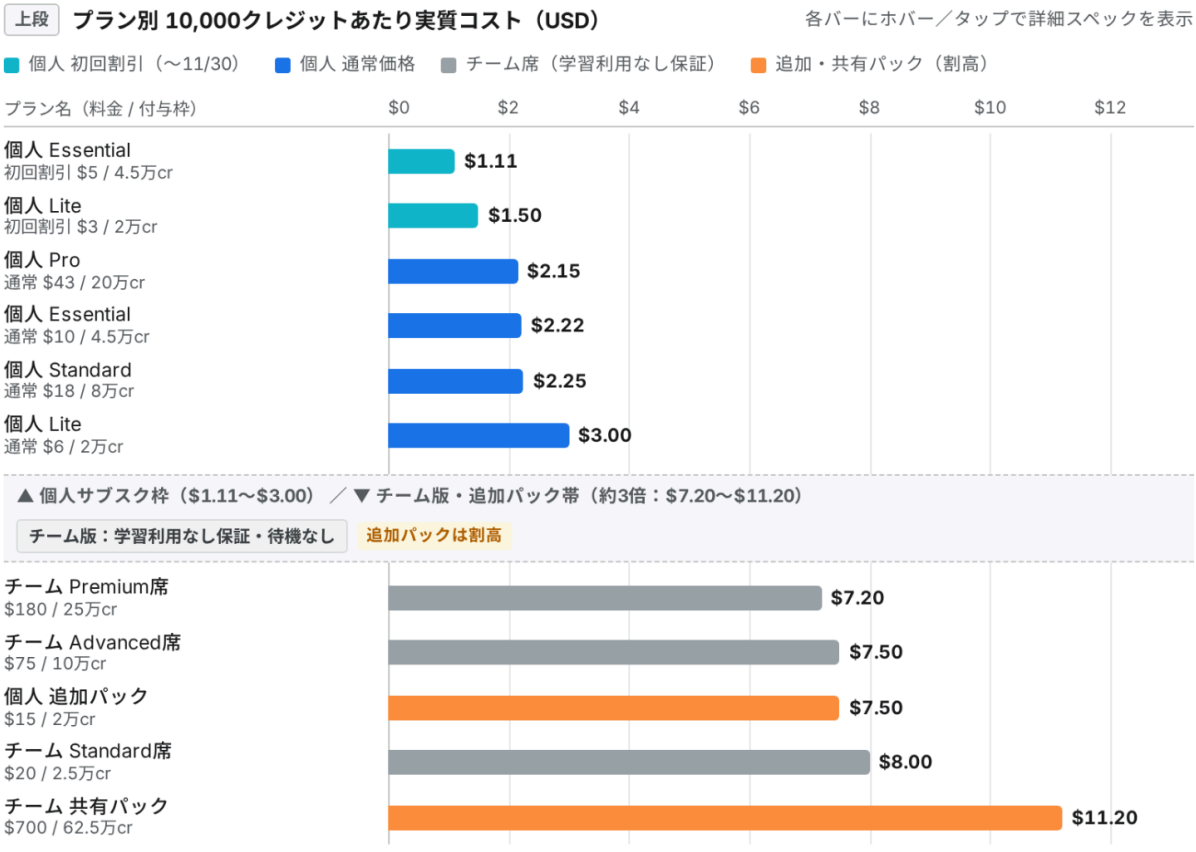
4. 利用ルート・Token Planの料金体系・データガバナンス

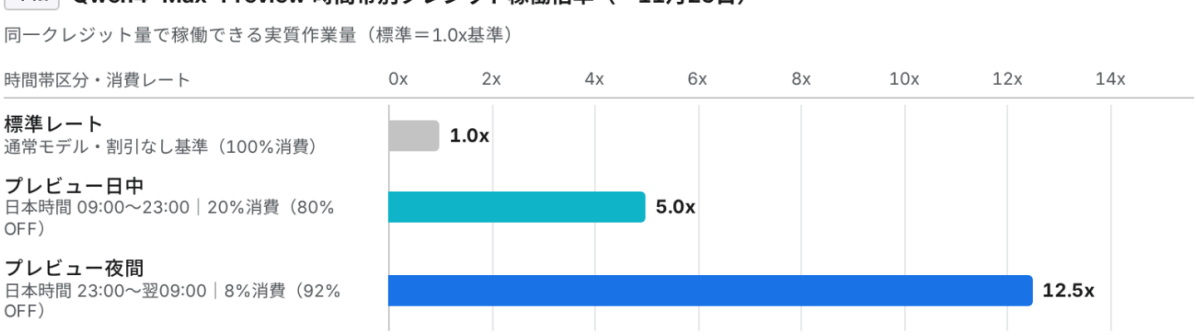
Qwen4-Max-Previewへの公式アクセス経路は現在6つのポート（Token Plan、Qwen Chat Web版、Qwenアプリ、Qoder、QwenWork／千问办公、QwenStudio）に分かれており[12]、利用目的とデータ保護要件に応じた選択が求められる[11]。

ブラウザから無料で試用する場合は、chat.qwen.ai（Qwen Chat／Qwen Studio）にアクセスしてサインインし、画面左上のモデル選択欄からQwen4-Max-Previewを選ぶだけで、テキスト・画像・動画を添付した対話が行える[8]。無料プランでの利用回数上限は公表されておらず、入力データは国外

サーバーで処理される前提となる[11]。一方、Claude Code、Cursor、Cline、Codex、Qwen Code、OpenCode、OpenClawなどのAIコーディング・エージェントツールからAPI経由で呼び出すには、国際版ではAlibaba Cloud Model StudioまたはQwenCloudのシンガポールリージョン（ap-southeast-1）で提供されている月額サブスクリプション「Token Plan」を契約し（中国本土向けの阿里雲百煉でも国内版Token Planを展開）、専用APIキー（sk-sp-で始まるキー）を取得する必要がある[6]。なお、Alibaba Cloud Model StudioとQwenCloudのToken Planはプラン名・価格が共通しているものの、決済システム・認証キー・クレジット残高は別個に管理されており互換性はない[43]。

Token Planは日本時間23時〜翌9時の夜間利用で実質12.5倍稼働しチーム版はデータ保護対価として単価が約3倍となる





上段は2026年11月9日のクレジット増量改定後の各プランにおける10,000クレジットあたりの米ドル単価（新規ユーザー割引は2026年11月30日までの初回契約者に適用）。個人版の上位プランほど単価が下がる一方、追加パック（$7.50）は割高なため枠不足時は日割りアップグレードが有利となる。チーム版は「学習利用なし保証」と「ピーク時待機なし」を含むため個人版の約3倍の単価設定となっている。下段は11月25日までのプレビュー割引による実質利用可能倍率を示す。

データ出典: Alibaba Cloud Model Studio、 QwenCloud Documentation、 eesel AI、 AI Revolution

クレジット経済性と11月9日の事前価格改定

AlibabaはQwen4-Max-Preview公開の2日前にあたる2026年11月9日にToken Plan個人版の価格・クォータ改定を実施し、Lite〜Standardの月間クレジット枠を74%〜78%増量（Lite：11,500→20,000、Essential：25,500→45,000、Standard：45,000→80,000）するとともに、最上位Proプランを月額68から 43へ37%値下げしつつ枠を180,000から200,000クレジットへ引き上げた[15]。Alibabaが公開しているqwen3.6-plusでのクレジット消費例をもとにした第三者試算（eesel.ai）では、1クレジットは従量課金の定価ベースで約

0.0025 USD相当（$1あたり約400クレジット）**と算出されており、通常月額$6（2026年11月30日23:59:59 UTC+8までの初回契約者は

3）のLiteプラン（20,000クレジット）で約50相当、月額 43のProプラン（200,000クレジット）では約$500相当の定価利用枠に相当するとされる[6]。

さらに11月25日24時（UTC+8）までのプレビュー期間中は、Qwen4-Max-Previewのクレジット消費量が日中は標準レートの20%（80%オフ＝通常の5倍稼働）に設定され、北京時間22時00分〜翌08時00分（日本時間23時00分〜翌午前9時00分）の夜間帯には追加で60%オフとなる標準レートの8%（92%オフ＝通常の12.5倍稼働）が適用される[5]。日本の開発者にとっては深夜から早朝にかけての作業時間帯がこの8%消費枠に重なるため、初回割引のLite（$3）やEssential（

5）でも多くの検証タスクを実行できる。なお、月途中でクレジットを使い切った場合、20,000クレジットあたり

15の追加パック（Extra Bundle／Credit Pack、最大5個まで保持可）を購入すると単価がLite通常価格の2.5倍（$7.50 / 1万クレジット）となるため、残日数に応じた日割り差額（(新価格 - 旧価格) × 残日数 / 30）で即時クォータが加算される上位プランへのアップグレードを行う方が経済的である（ただし途中解約・返金・ダウングレードは不可であり、未使用クレジットの翌月繰越も行われない）[6]。また、単一のToken Plan契約でQwenシリーズ以外にもdeepseek-v4.1-flash、deepseek-v4-pro、glm-5.3、wan3.0-video、さらにチーム版ではkimi-k2.7-codeやMiniMax-M2.5を含む20種以上のモデルを共通クレジットで利用できる[6]。

企業導入におけるデータプライバシーと利用規約上の境界線

個人版とチーム版の間に設けられた「1クレジットあたり約3倍の単価差」は、そのまま法人向けのガバナンス保証とインフラ品質の差に対応している[15]。

利用ルート / プラン区分	料金目安（USD）	学習データ利用（プライバシー）	許可される用途とインフラ特性	想定ユーザー層
Qwen Chat / Studio・アプリ (chat.qwen.ai)	無料[8]	国外サーバー処理・一般規約準拠[11]	ブラウザ・アプリ上での手動チャット（回数上限非公表）[11]	手軽に日本語・画像・コード生成を試したい一般・開発者[11]
Qoder / QwenWork	アプリ内課金（Qoderは80%割引報道）[22]	各ツール規約準拠（Token Plan紐付け可）[6]	専用IDE内での開発（Qoder）またはオフィス文書・PPT作成（QwenWork）[12]	環境構築不要のIDEや完成ドキュメント出力を求める実務者[41]

Token Plan 個人版（Lite / Essential / Standard / Pro）	月額 6〜43（初回Lite $3・Essential $5）[6]	標準サービス規約準拠（明示的な不学習保証なし・越境転送あり）[6]	対話型コーディングツール限定（同時1〜8エージェント、ピーク時待機発生の可能性あり）[6]	個人開発者、公開情報や検証用リポジトリでの技術評価[15]
Token Plan チーム版（Standard / Advanced / Premium）	1席あたり月額 $20 / $75 / $180[6]	会話データをモデル学習に使用しない明示的コミットメントあり[6]	対話型ツール限定・マルチテナント分離によりピーク時のキュー待機なし・席管理コンソール付[6]	社内コードや機密業務データを用いて検証を行う企業開発チーム[6]
サードパーティ **API**中継（CometAPI・AIHubMix等）	独自設定（例：CometAPI 入力 $1.6/出力$ 4.8 per M）[1]	中継事業者のポリシーに依存[1]	公式は「Token Plan専売・従量課金非対応」と明記しており、規約抵触による遮断リスクを伴う[6]	公式ドキュメント外の非正規経路（業務利用は非推奨）[6]

業務評価において特に留意すべきは以下の2点である。第一に、データ学習ポリシーの差異である。Token Planチーム版の公式ドキュメントには「会話データはモデル学習に使用されない（Conversation data is not used for model training）」と明記されているのに対し、個人版にはその明示的コミットメントがなく標準サービス規約に委ねられており、かつ推論はシンガポールリージョンを起点とするグローバル展開モード（国境を越えるデータ転送を伴う）で実行される[6]。社内機密コードで評価する組織にはチーム版が適している[15]。第二に、個人版・チーム版共通の「対話的利用限定」規約である。Token PlanのAPIキー（sk-sp-）は、契約者本人が所有する複数端末（自宅PCと職場PCなど）へ設定することは認められているが、他者とのキー共有（チーム版も1席1キー専有）や、サポートチャットボット、チケット分類、自動化スクリプト、自社アプリケーションのバックエンドAPIとしての非対話的・バッチ呼び出しは厳格に禁止されている[5]。なお、CometAPIやAIHubMix、Infronなどの一部サードパーティAPI中継サービスが独自の従量単価や無料枠でqwen4-max-previewの提供を謳っている例も見受けられるが[1]、Alibaba公式は「プレビュー期間中はToken Planのみで利用可能であり従量課金は非対応」と明言しているため、外部中継ルートへの依存はキー停止や提供中断のリスクを伴う[6]。

5. 実測パフォーマンスと多角的評価：公式ベンチマーク不在の中で見えた実力

Alibabaによる公式ベンチマーク表が存在しない中、Qwen4-Max-Previewの客観的な性能特性を判断する材料となるのが、公開直後の2026年11月11日〜15日にかけて各国メディア、AI検証ブロ

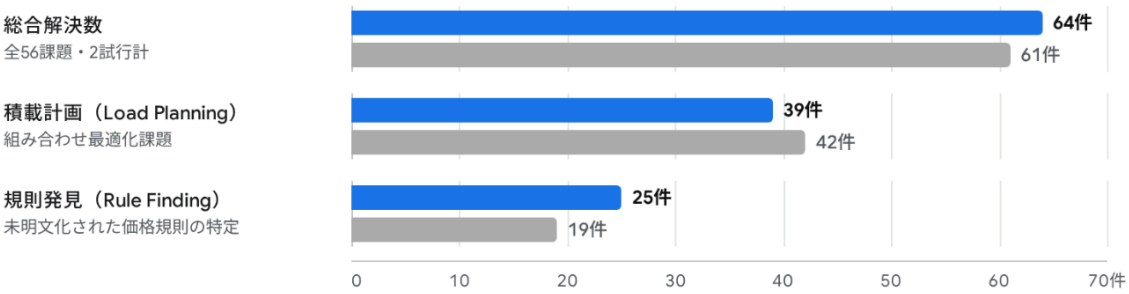
グ、および開発者コミュニティが実施したハンズオン検証と独自ベンチマークである[8]。

実測では帰納的ルール発見と複雑な実装完遂力で強みを示す一方、深い推論により実行時間は長大化する

サンドボックス型エージェント比較：課題解決数

Qwen4-Max-Preview　DeepSeek-V4.1-Flash

Hermes環境・ネット遮断・新規生成56課題×2試行（計112試行、上限25分/試行）における解決件数



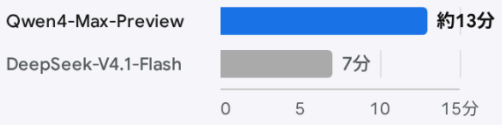
同ベンチマークにおける実行速度と時間切れ頻度

上限到達タイムアウト警告値

課題解決数が拮抗する一方、Qwen4-Max-Previewは約2倍の思考時間を要しタイムアウトが4倍発生

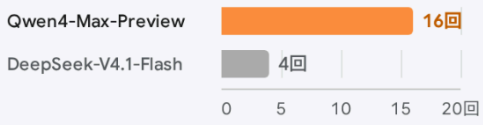
1課題あたりの平均所要時間

単位：分（短いほど高速）



25分上限到達タイムアウト数

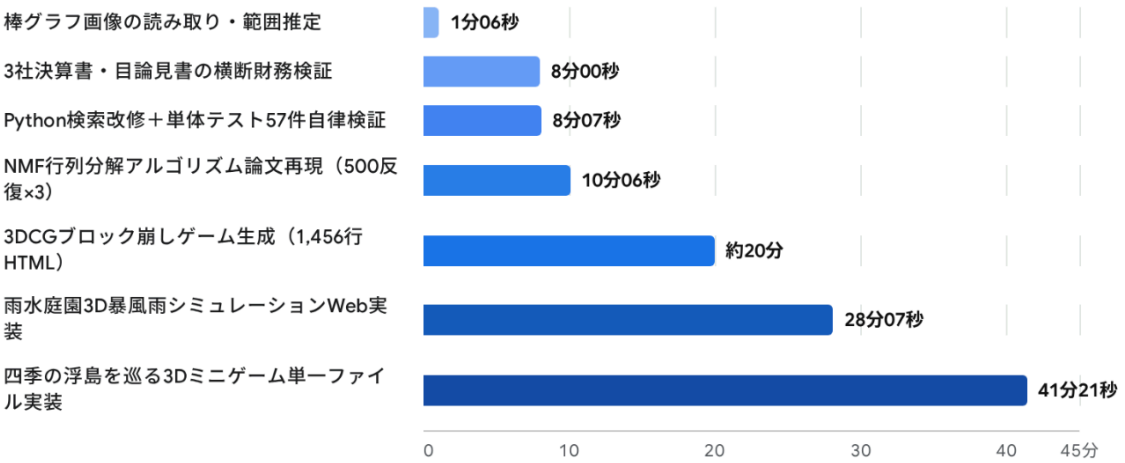
単位：回（全112試行中の時間切れ回数）



タスク複雑度別の実測ウォールクロック実行時間

各バーにホバー／タップで成果物詳細

12シナリオ検証およびハンズオンで記録されたQwen4-Max-Previewの単一タスク完了時間（分）



上段・中段はインターネット遮断サンドボックス環境（Hermesエージェント、新規生成56課題、制限時間25分/試行）での比較結果。総合解決数ではQwen4-Max-Preview（64件）とDeepSeek-V4.1-Flash（61件）が拮抗しつつ、未明文化された価格規則を特定する「ルール発見」でQwen4が優位に立った反面、1課題あたり平均約13分を要し16回の時間切れが発生した。下段は各種ハンズオン検証で記録された単一タスクあたりの実測生成・実行時間。

Data sources: Reddit r/Qwen_AI, Reddit r/DeepSeek, GIGAZINE, Sohu (搜狐科技)

1. 自律コーディングとエージェント遂行能力

Reddit（r/Qwen_AI）で公開された事前学習汚染排除型のサンドボックス検証（インターネット遮断下のHermesエージェント内で、記憶不可能な新規生成課題56問×各2試行＝計112試行、制限時間25分/試行で比較；単一ラン・2課題種別の小規模サンプルである点には留意が必要）では、Qwen4-Max-Previewは計64件を解決し、高速モデルであるDeepSeek-V4.1-Flashの61件と総合スコアで並んだ[18]。内訳を見ると、「積載計画（Load Planning）」ではDeepSeekが42対39でやや上回った一方、数百件の請求データから未明文化の価格ルールを帰納的に特定して新規30件を1セント単位で計算する「規則発見（Rule Finding）」ではQwen4-Max-Previewが**25**対**19**で優位に立ち、両モデルとも0勝24敗となった最難関3段階の手前でもDeepSeekより1段階上まで到達した[18]。ただし1課題あたりの平均所要時間はDeepSeekの7分に対してQwen4-Max-Previewは約13分を要し、25分上限によるタイムアウトも4回対16回と4倍発生している[18]。また、別の開発者による4課題（コードレビュー・実装・法務・技術文書）の独自比較では、xhigh設定のQwen4-Max-PreviewはAnthropicのClaude Opus 5.5（Medium設定）と同等（Opus 5.5 Lowよりは明確に上、ただしトークン消費量と所要時間は約2倍）と報告されており、eesel.aiの初期レビューでもxhigh思考時の体感速度はOpenAIのGPT-6.1 Terraよりやや遅いと評されている[5]。

ソフトウェアエンジニアリングにおける自律完遂力は、中国Sohu Techが実施した12シナリオ検証のPython工単（チケット）検索サービス改修テスト（所要8分07秒）で確認されている[12]。フィルタリング・ソート・ページネーション・入力検証に不具合を抱えた既存コードに対し、Qwen4-Max-Previewは既存の単体テスト9件を実行して2件の失敗を再現したのち、query.pyの処理ロジックをリファクタリングし、自ら48件の新規単体テストを追加（途中でURLエンコーディングと中国語検索仕様に起因する3件の失敗を自律検知して自己修正）、全57件のテスト通過と旧コードの失敗による逆検証を経て、13回のライブHTTPリクエスト（200応答7回、400応答4回、404応答2回）で動作を実証し、元データのハッシュ値を維持したまま改修パッケージを出力した[12]。フロントエンド・3D実装でも、Simon Willison氏の定番課題「自転車に乗るペリカンのSVG」でペダルに足を乗せた構造と再生速度調整UI付きWebページを一発生成したほか、外部依存ゼロの単一オフラインファイルで動作する暴風雨時の雨水庭園3D断面シミュレーション（所要28分07秒、ピーク水深が152.5mmから53.75mmへ64.8%低下する61ステップの状態スナップショットを追跡）や、四季の浮島を機関車で巡る3Dミニゲーム（所要41分21秒）を完遂している[12]。

2. 長文脈検索・財務読解と「ハルシネーション抑制力」

Sohu TechやMindStudioの検証で共通して高く評価されているのが、不確実な情報に対する慎重さとハルシネーション（事実捏造）の抑制である[12]。Sohu TechがApple（FY2026）、Costco（FY2026）、およびAnthropic（11月10日第2ドラフト目論見書）の3社資料を横断読解させた検証（所要8分00秒）では、抽出対象の主要財務数値6項目すべてが一致し、Anthropicの営業損失（12.6億）と転換社債の公正価値変動を含む純損失（1,040.39億）の乖離原因も正確に切り分けた[12]。さらに「2027年の売上高ガイダンスは？」という引っかけ質問に対して「3社とも未開示」と回答し、「Anthropicの時価総額は？」という問いにも「70〜78は想定株価レンジであり、本文中の2兆ドルは世界の医薬品支出を指す」と文脈混同のトラップを正確に回避した[12]。

また、約500KB（顧客チケット1,600件分）の生ログを用いた長文脈「Needle-in-a-Haystack」検証では、741行離れた2つのチケットに分割された同一顧客の返金総額（1,386人民元）の特定や、「華東地域＋物流遅延」の該当件数集計（62件、プログラム集計と一致）など全5問に正解している[12]。一部のみ数値ラベルが付いた棒グラフ画像の読み取り（所要1分06秒）や、乱雑な売上明細44行のクレ

ンジング（スペース違いの重複や「万元／元」の単位混在を整理し、動的に再計算されるExcel数式入りの35行シート・合計215,700人民元を出力）においても、未記載の数値を捏造せずレンジ推定や「会議内で未決定」の明記を守る堅実な挙動が確認されている[12]。

3. 視覚知能（Visual Intelligence）と実世界物理・法務推論の統合

前世代Qwen3やQwen3.8で課題とされていた視覚認識の細部精度も改善が報告されている[17]。MindStudioのハンズオン検証では、後部座席の子どもを写した画像から「シートベルトが肩ではなく脇の下を通っている」という細部（Qwen3が見落とした視覚特徴）を検出し、さらにアメリカ、オーストラリア、インド、ポーランド、チリなど9カ国について「条文上の規定」と「実際の警察による罰金執行の可能性（オーストラリアのみ実質的な罰金リスクが高く、対象は子どもではなく運転手である点など）」を区別して推論した[17]。また、オーストラリア南東部の衛星画像から雨陰効果を分析する課題（所要10分超）では、当て推量を避けて自らPythonコードを作成・実行し、12組の町ラベル間のピクセル距離から縮尺（約1km/pixel）を算出、途中の町名取り違えを思考過程で自己修正したうえで、大気物理学の断熱減率（1,000m上昇で約5.5°C低下、下降で約9.8°C上昇）を適用して乾燥側の約4°Cの昇温を導き出し、植生境界の不確実性（約320km±50km）を明示した[17]。

4. 日本語での利用感と思考言語の挙動

Alibabaは日本語単体の定量評価スコアを公表していないものの、対応言語は日本語・中国語・英語・アラビア語など40言語以上に及ぶ[11]。MindStudioの多言語検証でも82言語すべてで重複なく現地の文化的連帯概念（ズールー語の*Ubuntu*やインドネシア語の*Gotong Royong*など）を踏まえた政党名を創案している[17]。日本のGIGAZINEによる実機レビュー（chat.qwen.ai無料プラン使用）では、「Windows版Chromeで実行可能なブロック崩しゲームを3DCGと豪華なエフェクト付きで単一HTMLファイルとして作成して」という日本語プロンプトに対し、画面上の思考プロセス（**reasoning_content**）は中国語で展開されつつも日本語の指示を的確に反映し、約20分かけて1,456行のHTMLファイルを生成した[8]。初回出力では読み込み画面から進めない不具合が生じたものの、日本語で「読み込み画面から先に進めない」と1回フィードバックしただけで正常にプレイ可能な修正版を生成しており、対話による反復改善のしやすさが確認されている[8]。

5. 実測で判明した弱点と失敗パターン

一方で、上記の各検証からは明確な限界も報告されている。Sohu Techの非負値行列因子分解（NMF）論文再現テスト（所要10分06秒、48×36行列・3種の初期化×500反復）では、数値計算自体は参照データと一致したものの、最終500ステップ目でも目的関数の誤差がまだ減少中であるにもかかわらず「連続的な減少は局所最適解への収束を証明している」と結論づける定性的な過剰解釈を犯した[12]。また、情報集約タスクにおける軽微な抜け漏れや推定誤差も見られ、平頭哥（T-Head）スピンオフ上場の編集可能PPT生成では20個の歴史的事実を維持しつつも製品一覧表でスマートNIC製品ライン（磐脉 / Panmai）1つを落とし、Anthropicの46KB対話型HTMLダッシュボードでは「12ヶ月以内に満期を迎える約70.7億ドルの債務」という詳細1点を書き漏らし、未ラベル棒グラフのWearables部門の上限推定でも実数値より6億ドル低い値を出した[12]。加えてMindStudioの検証では、HTML/CSS/JSによる屋台アニメーションでカメラ角度が変化した際に非定型なカスタムオブジェクト（回転串の丸焼き動物）の描画が崩れた点や、82言語の命名課題で「名前のみを返せ」という指示に対して短い解説文を添えてしまう軽微な指示逸脱（Instruction Drift）が指摘されている[17]。

6. 開発者向けAPI実装ガイドと既知の落とし穴・不具合

Token Planを契約して開発ツール（Claude Code、Cursor、Codex、Qwen Code、OpenClaw等）や

検証環境からqwen4-max-preview（またはエイリアスqwen4.0-max-preview；エージェント側のモデルID指定は完全一致のホワイトリスト判定であり、請求書上は一律qwen4.0-max-previewと記録される）を呼び出す場合、専用Base URLはAlibaba Cloud Model Studio契約時はhttps://token-plan.ap-southeast-1.maas.aliyuncs.com（OpenAI互換パスは/compatible-mode/v1）、QwenCloud契約時はhttps://token-plan.maas.qwencloudapi.com（Anthropic互換パスは/apps/anthropic）を使用する[3]。公式のBailian CLI（bailian.aliyun.com/cli/install.sh）のbl config agentコマンドを利用すれば、主要エージェントツール向けの接続設定を自動構成することも可能である[6]。

実装およびツール運用にあたっては、従来のQwen3.8系や標準OpenAI APIとは異なる以下の6項目の技術的仕様・初期不具合を事前に確認しておく必要がある。

1. **max_tokens**仕様変更による出力途中切断の回避：OpenAI Chat Completions互換APIにおいて、max_tokensはmax_completion_tokensと同義として扱われ、思考トークン（**reasoning_content**）と最終回答（**content**）の合算値を制限する仕様に変更された[5]。既定のxhigh設定では思考だけで大量のトークンを消費するため、最終回答サイズのみを想定した小さな上限値を設定すると回答が始まる前に出力が切断される[5]。十分な上限（最大393,216）を確保するか、特に構造化出力（JSON）利用時は途中切断によるパース破損を防ぐためmax_tokens自体を未指定にすることが公式に推奨されている[6]。
2. マルチターン会話における**reasoning_content**の分離返送：Chat Completionsではpreserve_thinkingがデフォルトで有効化されている[3]。複数ターンの会話やFunction Callingの2回目以降の呼び出しでは、過去のassistantメッセージの思考内容を独立した**reasoning_content**フィールドに格納して返送する必要があり、contentフィールドに連結して送信することはできない（非対応のSDKやツールではマルチターン処理が失敗する要因となる）[3]。
3. 推論パラメータの排他制約と**tool_choice**の制限：思考強度を指定するreasoning_effort（"low", "medium", "xhigh"；Python版OpenAI SDKではextra_body経由で指定）と、トークン上限を直接指定するthinking_budgetを同時に設定することはできず、併用するとエラーが返される[3]。また、思考モードが有効な状態では特定関数の呼び出しを強制するtool_choice（"auto"と"none"のみ対応、"required"は非対応）が利用できないため、特定関数を強制する場合は先にenable_thinking: falseを設定する必要がある（Token Planの自動モデル選択autoもtool_choiceやthinking_budgetには非対応）[6]。
4. サーバー側中国語ルール（**routify**）の漏洩と自己プロンプトインジェクション誤検知バグ：GitHub（QwenLM/qwen-code Issues）およびReddit（r/Qwen_AI, r/LocalLLaMA）において、サーバー側で注入されている一時プロキシURL規制に関する中国語ルール（routify等临时代理地址；任务明确需要且已有真实链接时除外…）がコンテキスト内に漏洩し、qwen4-max-preview自身が思考トレース内でそれを「プロンプトインジェクションの疑い」として検知・言及してしまう事象が複数報告されている[20]。クライアント側の不具合ではなくプラットフォーム側のルール注入に起因する挙動として留意が必要である[20]。
5. 入力データ検査エラー（**data_inspection_failed**）とリトライループ対策：Alibaba Cloud百煉（Model Studio）のコンテンツ検査機構により、特定の文字列を含む入力がHTTP 400エラー（{"code": "data_inspection_failed", "message": "Input text data may contain..."}）で拒否される場合がある[56]。Qwen Code（Issue #2905）等の一部クライアント環境では、同エラー発生時に同一コンテキストを繰り返し再送してリトライループに陥る事例が報告されているため、400

系エラー時の即時停止やコンテキスト整理が必要となる[56]。

6. 長大レイテンシ対策と動的レート制限（**429**）への対応：深い推論に伴う応答遅延（外部計測サイトInfronの無料枠では24時間平均TTFTが39.61秒、AIHubMix計測のスループットは約28〜62 tok/s）に対処するため[46]、公式ドキュメントはストリーミング出力（**stream: true**）の利用と、クライアント側のタイムアウト設定を**180**秒以上へ延長すること（単純な質疑応答ではenable_thinking: falseで出力時間を60%〜75%短縮可能）を推奨している[6]。また、Token Planのレート制限（TPM/RPM）は全体負荷に応じて動的に調整され個別引き上げは不可であるため、429 Requests rate limit exceeded発生時は並列エージェント数やコンテキスト量を抑えて約1分間待機してから再試行する運用が求められる[59]。

7. 総合結論：今すぐ試すべき用途と本番導入に向けた意思決定指針

2026年11月15日現在、公式プレビュー終了（日本時間11月26日午前1時00分）まで残り約10日間というスケジュールを踏まえると、Qwen4-Max-Previewに対する合理的なアプローチは、本番システムへの直接依存を避けつつ、最大**92%**オフのクレジット優遇が適用されるプレビュー期間を活用して、自社の難度の高い開発・分析課題における次世代モデルの到達点を検証することにある[1]。

まず期間中に優先して試すべき用途としては、無料のchat.qwen.ai、あるいは社内コードを扱うならデータ不学習保証を備えたToken Planチーム版（公開データでの個人検証なら初回3〜5の個人版と日本時間23時〜翌9時の8%消費枠の組み合わせ）を通じた、複数ファイルにまたがる既存コードのテスト駆動リファクタリング、最大39.3万トークンの出力枠を活かした単一パスでの大規模フロントエンド・3Dプロトタイピング、数十万トークン規模のログや複数決算書の突合分析、および高精細な画像・図表と専門知識を組み合わせた多段階推論タスクの評価が挙げられる[6]。

反対に、自社Webサービスや社内自動化パイプラインのバックエンドAPIとして稼働させる本番用途には、利用規約上の禁止条項および11月25日のモデル退役という双方の理由からQwen4-Max-Previewは使用できない[5]。Qwenファミリーでの本番組み込みには従量課金API（2/6 per 1M）とオープンウェイトが提供されているQwen3.8-Maxまたは高効率なQwen3.8-Flash-Nextを継続採用すべきであり[27]、低レイテンシが求められるエージェント実行にはDeepSeek-V4.1-Flash[18]、最高水準の本番マルチエージェント開発環境には同時期に一般提供されているClaude Fable 5.2、Claude Opus 5.5、Gemini 4 Carbonなどの正式リリースモデルを選択するのが堅実である[13]。

11月25日のプレビュー終了後には、今回の公開テストの知見を反映した正式版Qwen4-Max（従量課金API対応）や、セルフホスト・ローカル運用向けとして期待されるQwen4-27BおよびQwen4-Flashの展開が見込まれる[14]。ただし現時点では正式版のリリース時期、トークン単価、ライセンス条件、および27BモデルのVRAM要件はいずれも未確定であるため、推測に基づくGPUハードウェアの先行調達やプレビュー用モデルID（qwen4-max-preview）の固定的な組み込みは控え、max_completion_tokens相当の共有出力バジェット管理とreasoning_contentフィールドの分離保持に対応した柔軟なクライアント実装を整えておくことが、正式版公開時に円滑に移行するための最善の準備となる[7]。

引用文献

1. Qwen4-Max-Preview をリリース - CometAPI, https://www.cometapi.com/ja/qwen4-max-preview-released/

2. 次世代AI「Qwen4-Max-Preview」を公開 一度に参照できる情報量は, https://realsound.jp/tech/2026/11/post-2582494.html
3. Alibaba Cloud Model Studio:qwen4-max-preview, https://www.alibabacloud.com/help/ja/model-studio/qwen4-max-preview
4. Token Plan (Personal Edition) - Alibaba Cloud Model Studio, https://www.alibabacloud.com/help/ja/model-studio/token-plan-personal-overview
5. https://www.eesel.ai/blog/qwen-4-max-review
6. https://www.alibabacloud.com/help/en/model-studio/qwen4-max-preview
7. https://docs.qwencloud.com/developer-guides/getting-started/latest-model
8. 「Qwen4-Max-Preview」が無料で使用可能になったのでゲームを, https://gigazine.net/news/20261112-qwen4-preview/
9. Alibaba、次世代AI「Qwen4-Max」をプレビュー公開。100万, https://news.mynavi.jp/article/20261112-5307569/
10. Qwen4-Maxがプレビュー提供開始、Alibabaは月3ドルから, https://pasqualepillitteri.it/ja/news/31979/qwen4-max-preview-token-plan-alibaba-ja
11. Qwen4とは？Qwen4-Maxプレビューの使い方・料金・提供期限を, https://ai-revolution.co.jp/media/what-is-qwen-4/
12. https://m.sohu.com/a/1101908446_122074763?scm=10001.325_13-325_13.0.0-0-0-0-0.5_1334
13. Alibaba Opens the Qwen4-Max Preview: 1M Context, Up to 393K, https://aicoder.com/news/news-20261111-alibaba-qwen4-max-preview-1m-context-393k-output-multimodal-token-plan-only
14. Release Date, Qwen4-Max Preview, and What's Confirmed (2026), https://www.yottalabs.ai/post/qwen-4-release-date-what-is-known-how-to-prepare-2026
15. Qwen 4 Max pricing (2026): Token Plan credits, plans and real cost, https://www.eesel.ai/blog/qwen-4-max-pricing
16. https://www.alibabacloud.com/help/en/model-studio/token-plan-permissions-and-usage
17. Qwen4 Max Hands-On: Testing Alibaba's New Model on Real Tasks, https://www.mindstudio.ai/blog/qwen-4-max-hands-on-tests
18. Qwen4 Max Preview is performing at DeepSeek V4.1 Flash levels, https://www.reddit.com/r/Qwen_AI/comments/1xus44s/qwen4_max_preview_is_performing_at_deepseek_v41/
19. Observations from seeing Qwen 4 max's outputs : r/LocalLLaMA, https://www.reddit.com/r/LocalLLaMA/comments/1xtt3rb/observations_from_seeing_qwen_4_maxs_outputs/
20. Issues · QwenLM/qwen-code - GitHub, https://github.com/QwenLM/qwen-code/issues
21. Qwen4のMax・Plus・Flash・27BとQoduck：判明していることと, https://zenn.dev/puchifan/articles/c60ec0d8d36c14
22. AI Realtime News — AI for Developing Countries | AIFOD, https://af.net/realtime/?country=United+States&country_page=5&news_page=3

23. gpt-oss-120b vs Qwen4-Max Preview — Benchmarks Compared, https://aireleasetracker.com/compare/openai/gpt-oss-120b/qwen/qwen4-max-preview
24. Qwen4 Max Model Specs & Sources (November 2026) | BenchLM.ai, https://benchlm.ai/models/qwen4-max
25. Qwen 4 Max PreviewはStealthmodelsによって評価されました - Reddit, https://www.reddit.com/r/Qwen_AI/comments/1xu4n0j/qwen_4_max_preview_has_been_evaluated_by/?tl=ja
26. Release Date, Expected Sizes and What Changes from Qwen 3.8, https://www.betterclaw.io/blog/qwen-4-vs-qwen-3-8
27. https://www.yottalabs.ai/post/qwen-4-max-specs-benchmarks-pricing-how-to-access
28. https://www.hunteralphahub.com/qwen4-max-preview
29. Model lifecycle and updates - 阿里云文档 - Alibaba Cloud, https://help.aliyun.com/en/model-studio/newly-released-models
30. Qwen3.8-Flash-Next: A Deep-Dive into Alibaba Qwen4 Architecture, https://local-ai-zone.github.io/blog/qwen3-8-flash-next-deep-dive.html
31. Qwen3.8-Flash-Next: Alibaba's Qwen4 Architecture Preview | Enera, https://www.eneralabs.com/blog/qwen38-flash-next-qwen4-architecture-enterprise-2026/
32. Qwen3.8-Flash-Next vs Qwen3.8-Max: 6Bアクティブで十分か？, https://www.orcarouter.ai/ja/blog/qwen-3-8-next-vs-qwen-3-8-max
33. Qwen3.8-Max-0902 vs Qwen4-Max Preview - AI Release Tracker, https://aireleasetracker.com/compare/qwen/qwen3.8-max-0902/qwen/qwen4-max-preview
34. Qwen3.8とは？4モデルの違い・料金・ローカルでの動かし方, https://lifrell-tech.com/1856/
35. Use deep thinking models via API - Alibaba Cloud Model Studio, https://www.alibabacloud.com/help/en/model-studio/deep-thinking
36. Qwen3.8-Max Open Weights: First Max-Class Qwen You ... - LLM Stats, https://llm-stats.com/blog/research/qwen3-8-max-open-weights
37. The Best Open Source LLMs (November 2026): Ranked by ... - Morph, https://www.morphllm.com/best-open-source-llm
38. Qwen3.8-Flash-Next — open 125B MoE with 6B... - AI/TLDR, https://ai-tldr.dev/releases/qwen-3-8-flash-next/
39. Официальный Qwen4-Max уже доступен в предварительной, https://www.reddit.com/r/Qwen_AI/comments/1xt5skh/official_qwen4max_now_in_preview/?tl=ru
40. Qwen 4 Deep Dive: Architecture and Benchmarks - Kie.ai, https://kie.ai/blog/qwen-4-architecture-shift-open-weight-engineering
41. Qwen4-Maxの使い方：今すぐ利用できるすべての方法 - Apidog, https://apidog.com/jp/blog/how-to-use-qwen-4-max/
42. GitHub - aipoch/open-science, https://github.com/aipoch/open-science
43. Qwen Pricing 2026: Free Chat, Subscription Plans & API Costs, https://try-qwen-ai.com/pricing/

44. Token Plan | QwenCloud, https://www.qwencloud.com/pricing/token-plan
45. Token Plan (Team Edition) - Alibaba Cloud Model Studio, https://www.alibabacloud.com/help/en/model-studio/token-plan-team-overview
46. Coding Qwen4 Max Preview vs Kimi K3.1 Pro - AI Model Comparison, https://aihubmix.com/compare/coding-qwen4-max-preview/kimi-k3.1-pro
47. Qwen4 Max Preview:free — API Pricing, Providers & Specs - Infron AI, https://infron.ai/models/qwen/qwen4-max-preview:free
48. Qwen4 Max Preview is performing at DeepSeek V4.1 Flash levels, https://www.reddit.com/r/DeepSeek/comments/1xurzoj/qwen4_max_preview_is_performing_at_deepseek_v41/
49. 生成AIニュース | AI法務ラボ - LegalAgent, https://legalagent.co.jp/legal-lab/ai-news/
50. 「Qwen4-Max-Preview」が無料で使用可能になったのでゲームを, https://news.livedoor.com/article/detail/32797938/
51. Claude Code - QwenCloud, https://docs.qwencloud.com/developer-guides/clients-and-developer-tools/claude-code
52. Qwen4-Max-Preview の落とし穴、300トークンの上限で返ってきた, https://note.com/kuma_aihubmix/n/n13a52014faf2
53. Qwen 4 Max Preview injecting routify suppression (in Chinese), https://www.reddit.com/r/Qwen_AI/comments/1xuwdeu/qwen_4_max_preview_injecting_routify_suppression/
54. guardrail in qwen4-max-preview thinking trace : r/LocalLLaMA - Reddit, https://www.reddit.com/r/LocalLLaMA/comments/1xunsz9/guardrail_in_qwen4maxpreview_thinking_trace/
55. OpenAI compatible - Chat - 阿里云文档 - Alibaba Cloud, https://help.aliyun.com/en/model-studio/qwen-api-via-openai-chat-completions
56. 新模型用不了· Issue #190 · kingcharleslzy-ai/dsh-purge - GitHub, https://github.com/kingcharleslzy-ai/dsh-purge/issues/190
57. API Error: 400 <400> InternalError.Algo.DataInspectionFailed: Input, https://github.com/QwenLM/qwen-code/issues/2905
58. Coding Qwen4 Max Preview API - $0.353/M In, $1.059/M Out, https://aihubmix.com/model/coding-qwen4-max-preview
59. Alibaba Cloud Model Studio:Token Plan FAQ, https://www.alibabacloud.com/help/en/model-studio/token-plan-faq
60. FAQ - Alibaba Cloud Model Studio, http://docs.modelstudio.console.alibabacloud.com/en/model-studio/token-plan-personal-faq