Qwen4 Max 正式リリース調査報告

内容 性能評価 利用者の評判 知財実務への示唆

調査基準日　2026 年 11 月 16 日 日本時間
本文中の引用番号は文末の参考文献に対応する。

GPT-6 Astra

1 調査の結論と範囲

Qwen4-Max は 2026 年 11 月 16 日に正式リリースされた。Alibaba の公式発表と同日更新の API 仕様書で確認できる。11 月 11 日公開の Qwen4-Max-Preview とは別のモデル ID で、正式版の従量課金 API が提供されている。［1］［2］

本報告の評価は、長時間の専門業務や成果物作成で前世代から進歩し、料金も下がった有力モデルである一方、総合性能、処理速度、実際の費用で最新の米国製モデルを全面的に上回ったわけではない、というものである。

・前世代からの改善は明確である。第三者の総合指数と評価タスク費用の双方で改善が確認できる。［9］［10］［11］

・強みは業務によって異なる。Web 開発、専門資料を用いた成果物作成、法務課題には好成績が見られる。［14］［15］［16］

・導入判断では推論設定、総費用、利用経路、データの保存・推論場所を確認する必要がある。［4］［23］［24］［25］

調査対象は公式発表、開発者ドキュメント、第三者ベンチマーク、実使用レビュー、公開コミュニティの反応である。正式公開当日のため、長期運用の評判はまだ限られる。プレビュー版の体験談を正式版の評価と区別した。本調査では独自の API 実測は行っていない。順位・料金・仕様は調査時点の掲載値であり、随時更新され得る。

2 正式版の仕様とプレビュー版との差

表 1　正式版の基本仕様

項目	確認できた内容
開発元	Alibaba Qwen チーム
正式版モデル ID	qwen4-max
基本構造	新しい Qwen4 アーキテクチャによる MoE
基幹部分の規模	2.4 兆パラメータ
トークンごとの活性化規模	710 億パラメータ

項目	確認できた内容
コンテキスト長	100 万トークン
入力と出力	入力はテキスト・画像・動画、出力はテキスト
推論強度	low、high、xhigh。標準は high
外部処理	ツール呼び出し、Web 検索、構造化出力など
API 互換性	OpenAI Chat Completions／Responses、Anthropic Messages、DashScope
重み公開	Max 本体は現時点で非公開。小規模 Qwen4 モデルの公開を予告

出典：パラメータ規模・公開方針は公式発表、入出力と API 機能は詳細仕様による。画像・動画を理解する機能と、モデル単体で画像・動画を生成する機能は区別が必要である。［1］［2］［4］

表 2　プレビュー版と正式版の主な相違

項目	プレビュー版	正式版
提供開始	11 月 11 日	11 月 16 日
モデル ID	qwen4-max-preview	qwen4-max
API 利用	Token Plan 限定	従量課金 API、Token Plan
推論の標準設定	xhigh	high
推論強度	low／medium／xhigh	low／high／xhigh
思考と回答の合計上限	393,216 トークン	詳細仕様では 262,144 トークン
提供期限	日本時間 11 月 26 日午前 1 時に終了予定	確認した仕様書に終了日の記載なし

出典：公式プレビュー仕様と正式版の詳細仕様。プレビュー版の終了予定日は、正式版の発売延期を意味しない。両者は別モデルとして扱われる。標準の推論強度が異なるため、無設定で比較すると条件がそろわない。［3］［4］

公開当日の記載不一致として、QwenCloud の商品ページには最大出力 393K という表示がある一方、詳細な開発者ドキュメントでは非思考時 131,072、思考と回答の合計 262,144 とされている。実装では詳細仕様を基準に、利用するエンドポイントで確認することが妥当である。バッチ処理とファインチューニングも商品ページの一般的な機能紹介だけでは判断できず、正式版の詳細仕様では非対応である。［4］［5］

3 アーキテクチャと実務機能

技術的な中心は規模の拡大より、計算の効率化にある。前世代 Qwen3.8-Max も 2.4 兆パラメータだが、活性化規模は 950 億だった。Qwen4-Max では 710 億へ減り、約 25%少ない活性化規模となる。［1］［6］

Alibaba 研究院の発表転載は、次の設計変更を説明している。［7］

・Gated DeltaNet と Qwen Sparse Attention を組み合わせ、履歴の圧縮と関連文脈の選択的な参照を行う。

・Gated Residual により、複数の残差経路とゲートを用いて層をまたぐ情報伝達を改善する。

・N-gram Embedding により、局所的なトークンの並びを参照する追加表現を導入する。基幹部分とは別に 3,280 億パラメータと記載される。

この説明からは、同程度の基幹規模で長文処理、知識表現、推論効率を改善する設計と読み取れる。総パラメータ数と、1 トークンを処理する際の活性化パラメータ数は異なる指標である。

専門ツールによる作業と成果物作成

公式資料では、ゲーム開発における実行と修正、製造業の図面理解・3D 化・加工シミュレーション、金融調査における根拠付き Excel と編集可能な PPT、法務資料の分類と照合、専門計算と実験を組み合わせた研究が紹介される。いずれも外部ツールや実行環境を含む利用例であり、モデル API だけで同じ環境が整うわけではない。［4］

Alibaba は加工プログラム作成の人手を約 475 分から 24 分へ減らした事例と、48 時間のエージェント実行と実験を通じた冷却液分子の探索を紹介している。提供元が選んだ実演・事例であり、一般的な製造現場や研究課題で同じ改善率が再現することの証明ではない。［1］

途中成果を見ながら方向を修正する仕組み

Qwen チームが公開した Live Delivery プラグインは、利用可能な初版を作り、その後も版を更新し、利用者が追加指示、予算変更、停止を行える仕組みである。公開コードには成果物の版管理やフィードバック処理が示されている。モデル性能だけでなく、長時間作業の進め方を改善する取り組みと位置づけられる。［8］

4 メーカー公表ベンチマークの評価

Alibaba の公開比較図から主要項目を抜粋した。すべて高いほど良い指標だが、異なる行の数値同士は比較できない。また、全モデルを同一の推論予算・実行環境で測ったことまで、この比較図だけで保証されるわけではない。［7］

表 3　メーカー公表の比較値

ベンチマーク	主な対象	Qwen4 Max	Claude Opus 5.5	GPT 6.1 Sol
DeepSWE v1.1	ソフトウェア修正	75.7	74.2	75.2
Terminal-Bench 4.0	複雑な端末作業	48.5	59.6	56.1
ProgramBench	ソフトウェア再現	31.5	65.0	52.5
Agents’ Last Exam v1.1	長時間タスク	46.7	43.4	42.6
JobBench	職務タスク	68.5	75.4	66.2

ベンチマーク	主な対象	Qwen4 Max	Claude Opus 5.5	GPT 6.1 Sol
BioMysteryBench	生物情報学研究	85.0	81.9	84.4
Chartography	グラフの視覚理解	54.2	66.3	67.8

出典：Alibaba 研究院掲載の比較図。表中のモデル名表記は読みやすさのため改行している。［7］

DeepSWE では比較対象を上回るが、GPT-6.1 Sol との差は 0.5 ポイントで、圧倒的な差とは言えない。Terminal-Bench や ProgramBench では差が残る。長時間の専門業務や生物分野には強みが見られる一方、図表理解がすべての競合より優れているわけではない。

この比較図には Claude Fable 5.2 や Gemini 4 Carbon が含まれていない。一部で Opus 5.5 を上回ったことを、現在の最先端モデル全体を超えたと読むことはできない。

5 第三者評価による位置づけ

総合知能指数と評価タスク費用

Artificial Analysis の Intelligence Index v5.0.1 と、同評価におけるタスク当たり費用を確認した。推論設定は表中の設定であり、全モデルに同じ計算予算を与えた比較ではない。過去の記事に記載された旧版の指数とは、そのまま比較しない。［9］［10］［11］

表 4　第三者の総合指数と評価費用

モデルと評価設定	指数	費用 米ドル／課題
Claude Fable 5.2　Max／Default Fallback	65	26.54
Gemini 4 Carbon　High	60	5.48
Claude Opus 5.5　Max／Default Fallback	59	10.04
GPT-6 Astra　Max	57	6.24
GPT-6.1 Sol　Max	55	1.27
GPT-6.1 Terra　Max	51	0.96
Kimi K3.1 Pro	46	3.68
Qwen4-Max	43	3.27
Qwen3.8-Max　0902	34	10.02

出典：Artificial Analysis の 11 月 16 日確認値。モデル名に設定が明記されない行は、出典の表示に従った。［9］［10］［11］

Qwen の世代間では指数 34 から 43、評価タスク費用 10.02 ドルから 3.27 ドルへ改善し、費用は単純計算で約 67%減となる。一方、同じ評価では GPT-6.1 Sol や Terra が、より高いスコアをより低い費用で達成している。前世代より費用効率が改善したことと、市場全体で最も費用効率がよいことは分けて評価すべきである。

Qwen4-Max の個別ページでは、調査時点で処理時間が未掲載であり、総合的な速度優位も確認できない。［11］

実際のエージェント作業と Web 開発

Agent Arena では Qwen4-Max の xHigh 設定が 14 位。約 1 万 4,700 セッションを基にした暫定評価で、中国開発モデルでは掲載上の最上位だった。ネット改善指標は約プラス 7.5%だが、単純なタスク成功率 7.5%という意味ではない。［12］［13］

Qwen4-Max の xHigh は Claude Sonnet 5.5 の High 設定の 15 位より上だが、Sonnet 5.5 の Max 設定は 4 位である。Qwen4-Max の順位幅は 12〜19 位と表示され、近接モデルとの差には不確実性がある。「Sonnet 5.5 を上回った」という表現だけでは比較条件が抜け落ちる。［12］

Code Arena の Web 開発評価では 6 位、1750 点の暫定評価で、順位幅は 3〜9 位と表示される。Web 画面やアプリの試作では有力な選択肢と評価できる。［14］

文書、表計算、プレゼンテーションなどの知識労働を測る AA-Briefcase v1.1 では 8 位、1686 点だった。分析内容と見せ方などを合わせた評価であり、資料を読み込んで成果物を作る仕事に強みがあるとの見方を支える。［15］

6 法務ベンチマークと知財実務への読み替え

Harvey LAB-AA v1.1 は、24 の法律実務分野にわたる 120 の非公開課題で、資料から法務成果物を作らせる評価である。主指標では、全評価項目を満たしても重大な幻覚があれば得点を与えない。［16］

表 5　Harvey LAB-AA v1.1 の比較

モデル	全項目合格 重大幻覚なし	個別項目 合格率	重大幻覚数 課題当たり
Qwen4-Max	9.4%	91.2%	0.36
GPT-6 Astra　Max	8.6%	90.6%	0.03
Claude Fable 5.2 Max／Default Fallback	7.2%	91.0%	0.75
GPT-6.1 Sol　Max	6.9%	90.6%	0.09

出典：Artificial Analysis。Qwen4-Max は掲載順位 4 位。主指標の 9.4%は「法律回答の正答率 9.4%」ではなく、厳しい全項目合格条件に基づく指標である。［16］

資料処理や成果物作成を含む複合的な法務課題で好位置にある点は評価できる。ただし、重大な幻覚の少なさでは GPT-6 Astra や GPT-6.1 Sol が優位である。順位が高いことは、引用や法的結論を無確認で採用できることを意味しない。

また、この評価は日本の特許法、審査基準、クレーム解釈、FTO の精度を直接検証したものではない。知財実務への適用は、個別の課題で根拠の正確性を確認する必要がある。

7 料金と仕事全体の費用

表 6　100 万トークン当たりの掲載料金

課金対象	Qwen3.8-Max	Qwen4-Max	低下率
通常入力	2.00 ドル	1.50 ドル	25%
出力	6.00 ドル	4.50 ドル	25%
暗黙キャッシュ命中入力	0.25 ドル	0.075 ドル	70%

出典：QwenCloud のドル建て商品ページ。契約地域、提供範囲、追加ツールの料金は別に確認する。明示的キャッシュには作成料金もある。［5］［17］

Alibaba は自社の Work360 評価で、トークン消費とタスク費用の削減を示している。ただし「タスク費用約 6 割減」はその評価条件での結果であり、すべての利用で 6 割安くなるという意味ではない。［1］［7］

利用者は、思考を含む入出力トークン数、ツール呼び出しと再試行、完了までの待ち時間、人による修正と根拠確認の時間を合わせて測る必要がある。出力単価が安くても、思考が長くなれば費用は増える。第三者評価でタスク費用が必ずしも最安ではないことは、この違いを示している。［9］［11］

8 利用者レビューと公開直後の評判

日本語で具体的な作業過程を示すレビューを確認した。ただし、次の実使用例はいずれも正式版発売前のプレビュー版である。［18］［19］

表 7　プレビュー版の実使用レビュー

実施者と時期	課題	報告された結果
WEEL 11 月 11 日	整体院の LP 制作	約 17 分。配色・イラストの統一感を評価。前世代より時間がかかった。
WEEL 11 月 11 日	補助金調査	約 8 分。出典付き比較表と締切カレンダーを作成。内容の確認は必要。
WEEL 11 月 11 日	既存コード修正	約 12 分。失敗を修正し、36 件のテストに合格。
GIGAZINE 11 月 12 日	3D 表現付きゲーム	約 20 分で HTML 生成。初回は起動せず、追加指示 1 回でプレイ可能になった。

出典：各実施者のレビュー。WEEL は評価観点を書いた README やセキュリティ用の指示が検証環境にあったと明記しており、検証手順のすべてをモデルが自発的に発見したとは評価できない。［18］［19］

正式版発売後の Hacker News では、AA のスコアとタスク費用を取り上げ、安い単価でも総費用では競合に負ける場合があるという意見や、推論強度を変えて比較すべきだという議論が見られる。速度に関する投稿には前世代の体験も混在する。［20］

Reddit にも前世代からの進歩を評価する声と、主要なコーディング試験で最先端群に届かないことを残念がる声がある。これらは自己選択された少数の投稿であり、利用者全体の満足度調査ではない。［21］［22］

現時点の受け止め方としては、前進は認められるものの全面的な首位奪取ではなく、Web 制作や専門成果物で試す価値がある、という整理が妥当と考えられる。正式版の日本語業務性能、混雑時の速度、失敗率については、今後の検証が必要である。

9 企業利用の条件とデータ管理

Token Plan と従量課金 API は同じではない。公式利用条件では Token Plan はコーディング・エージェントツール内の対話利用向けであり、自動化スクリプト、独自アプリのバックエンド、非対話型の一括呼び出しに使えないとされる。毎晩の SDI 処理や大量公報の自動分析には、従量課金 API を前提に設計する必要がある。［23］

表 8　利用経路別に確認すべき条件

項目	公式資料から分かること
Token Plan 個人版	データ利用はサービス契約の条件に従う。
Token Plan チーム版	会話データをモデル学習に使わないと明記される。
Model Studio API の標準保持	原則として入出力を最大 30 日保持。法令・安全確保等による例外条件がある。
Zero Data Retention	対象企業、対象モデル、ワークスペース設定等の条件付き。自動的に全呼び出しへ適用されるわけではない。
東京リージョン	保存先と推論実行場所は別。Global では日本国外の計算ノードを利用し得る。

出典：Token Plan の利用方針、Model Studio のデータ保持方針、リージョンと推論範囲の説明。［23］［24］［25］

東京リージョンを選ぶだけで、推論も必ず日本国内になるとは限らない。Qwen4-Max の仕様では東京は Global 範囲で掲載され、Global の推論先には中国本土を含む利用可能なノードが含まれる。日本国内限定の提供可否はモデル別に確認する必要がある。［2］［25］

安全性について Alibaba は、コードの安全性確認、防御用途を重視した学習データ選別、危険用途に対する追加対策を説明している。ただし、この発表だけから独立した安全性検証まで完了したとは判断できない。［1］

10 日本語と知財実務での評価方針

今回確認した範囲では、正式版の日本語特許実務に特化した十分な第三者評価は見つからなかった。以下は、確認できた機能と評価結果を踏まえた本報告の応用判断であり、実証済みの性能を示すものではない。

表 9　知財業務における位置づけと検証項目

想定業務	現時点の位置づけ	検証すべき点
公報分類と SDI 一次評価	有力な検証候補	見逃し率、分類の一貫性、根拠段落
中国語特許と技術資料	比較評価する価値がある	数値、否定、条件、専門用語の訳し分け
複数公報の構成要件対比	長文と構造化出力を生かせる可能性	要件脱落、実施形態と請求項の混同
IP ランドスケープ資料	成果物作成能力と相性がよい可能性	集計の再現性、Excel と PPT の整合
契約と法務調査補助	法務試験が評価材料になる	引用、当事者、日付、義務の取り違え
未公開発明と社外秘資料	契約と処理場所を確定して判断	利用経路別のデータ条件
FTO と無効性の最終判断	補助分析として評価	原文・法的根拠との照合、重大誤り

100 万トークンの入力枠は魅力だが、大量の文献を入力できることと、すべての数値、例外、引用関係を正確に扱えることは別である。離れた段落の条件を組み合わせる課題、類似する実施例を区別する課題、根拠がない場合に回答を留保する課題で確認することが有効である。

ローカル AI との関係では、Max 本体の重みは現時点で公開されていない。小規模モデルの公開は予告されているが、サイズ、公開日、ライセンス、日本語能力、Max との性能差は公開後に個別に評価する必要がある。［1］

導入候補として比較する場合は、公開公報、契約、統合報告書などから 30 件程度の課題を選び、Qwen4-Max の high と xhigh を既存モデルと比較する。重大誤り、引用の正確性、修正時間、完了時間、総費用を共通指標として測ることを推奨する。公開結果が示す強みは、根拠付きの成果物をどこまで仕上げられるかという観点で確かめる価値がある。

参考文献

本文中の［番号］は以下の文献番号に対応する。すべて 2026 年 11 月 16 日閲覧。随時更新ページは同日の掲載内容を用いた。公開日等を確認できなかった資料には推定日を付さず、媒体や資料の性格を併記した。

［1］Alibaba Cloud Community. Alibaba Unveils Qwen4-Max New AI Teammate for Work Creativity and Discovery.

2026 年 11 月 16 日公開。閲覧日：2026 年 11 月 16 日。公式発表。提供開始、規模、実演、安全対策、小規模モデル公開方針。

https://www.alibabacloud.com/blog/alibaba-unveils-qwen4-max-new-ai-teammate-for-work-creativity-and-discovery_603783

［2］Alibaba Cloud. Qwen4-Max Latest Model.

2026 年 11 月 16 日更新。閲覧日：2026 年 11 月 16 日。正式版のモデル ID、入出力、API 互換性、地域別機能。

https://www.alibabacloud.com/help/en/model-studio/qwen4-max

［3］Alibaba Cloud. qwen4-max-preview.

2026 年 11 月 16 日更新。閲覧日：2026 年 11 月 16 日。プレビュー期間、出力上限、推論設定、Token Plan による提供条件。

https://www.alibabacloud.com/help/en/model-studio/qwen4-max-preview

［4］QwenCloud. Latest model Qwen4-Max.

公開日・更新日の明示を確認できず。閲覧日：2026 年 11 月 16 日。詳細仕様、推論設定、ツール、キャッシュ、利用例、地域と提供範囲。

https://docs.qwencloud.com/developer-guides/getting-started/latest-model

［5］QwenCloud. Qwen4-Max.

公開日・更新日の明示を確認できず。閲覧日：2026 年 11 月 16 日。商品ページ。ドル建て API 料金、キャッシュ料金、出力上限の表示。

https://www.qwencloud.com/models/qwen4-max

［6］Alibaba Group. Alibaba Unveils Qwen3.8-Max Its Largest and Most Capable Flagship Model to Date.

2026 年 8 月 3 日公開。閲覧日：2026 年 11 月 16 日。前世代の総規模 2.4 兆、活性化 950 億など。

https://www.alibabagroup.com/en-US/document-2021044032125272064

［7］阿里研究院. Qwen4-Max 正式上线.

2026 年 11 月 16 日 10 時 29 分掲載。閲覧日：2026 年 11 月 16 日。Alibaba 研究院による発表転載。アーキテクチャ、比較図、Work360、実演。比較表は同記事掲載画像を照合。

https://m.sohu.com/a/1102096678_384789

［8］Qwen Team. Qwen-Live-Delivery-Plugin.

公開日・更新日の明示を確認できず。閲覧日：2026 年 11 月 16 日。公式 GitHub リポジトリ。継続納品、利用者の介入、版管理の実装説明。

https://github.com/QwenLM/Qwen-Live-Delivery-Plugin

［9］Artificial Analysis. LLM Leaderboard Comparison of AI models.

随時更新。閲覧日：2026 年 11 月 16 日。Intelligence Index v5.0.1、設定別スコア、評価タスク当たり費用。

https://artificialanalysis.ai/leaderboards/models

［10］Artificial Analysis．Qwen3.8 Max 0902 Intelligence Performance and Price Analysis.
随時更新。閲覧日：2026 年 11 月 16 日。前世代の指数 34、タスク費用 10.02 ドル。
https://artificialanalysis.ai/models/qwen3-8-max

［11］Artificial Analysis．Qwen4 Max Intelligence Performance and Price Analysis.
随時更新。閲覧日：2026 年 11 月 16 日。正式版の指数 43、タスク費用 3.27 ドル、速度指標の掲載状況。
https://artificialanalysis.ai/models/qwen4-max

［12］Arena．Agent Arena Capabilities Overall.
随時更新。閲覧日：2026 年 11 月 16 日。エージェント実作業の暫定順位、設定、セッション数、順位幅。
https://arena.ai/leaderboard/agent

［13］Arena．Agent Arena Methodology.
公開日・更新日の明示を確認できず。閲覧日：2026 年 11 月 16 日。改善指標の定義と推定方法。単純な成功率とは異なる。
https://arena.ai/blog/agent-arena-methodology

［14］Arena．Code Arena WebDev Overall.
随時更新。閲覧日：2026 年 11 月 16 日。Web 開発評価の暫定順位、スコア、順位幅。
https://arena.ai/leaderboard/code/webdev

［15］Artificial Analysis．AA-Briefcase v1.1 Agentic Knowledge Work Benchmark.
随時更新。閲覧日：2026 年 11 月 16 日。知識労働の評価方法、分析・表現等の複合指標、順位。
https://artificialanalysis.ai/evaluations/aa-briefcase

［16］Artificial Analysis．Harvey LAB-AA v1.1 Benchmark.
随時更新。閲覧日：2026 年 11 月 16 日。120 課題・24 実務分野。全項目合格指標、個別項目合格率、重大幻覚数。
https://artificialanalysis.ai/evaluations/harvey-lab-aa

［17］QwenCloud．Qwen3.8-Max.
公開日・更新日の明示を確認できず。閲覧日：2026 年 11 月 16 日。前世代の入力・出力・暗黙キャッシュのドル建て料金。
https://www.qwencloud.com/models/qwen3.8-max

［18］WEEL．Qwen4-Max とは アリババ次世代フラッグシップのプレビュー版の特徴 料金 使い方を徹底解説.
2026 年 11 月 11 日公開・同日検証。閲覧日：2026 年 11 月 16 日。実使用レビュー。LP 制作、補助金調査、コード修正。検証環境の補助指示も明記。
https://weel.co.jp/media/tech/qwen4-max/

［19］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.
2026 年 11 月 12 日公開。閲覧日：2026 年 11 月 16 日。実使用レビュー。ゲーム生成、初回起動失敗、追加指示後の動作。
https://gigazine.net/news/20261112-qwen4-preview/

［20］Hacker News．Qwen4-Max Discussion.
2026 年 11 月 16 日の公開討論。閲覧日：2026 年 11 月 16 日。利用者・開発者による費用、推論設定、速度等の議論。統計的な満足度調査ではない。
https://news.ycombinator.com/item?id=50513215

［21］Reddit r/Qwen_AI. Qwen4-Max is here How competitive is it.
2026 年 11 月 16 日投稿。閲覧日：2026 年 11 月 16 日。公開直後の反応。検索取得で確認した範囲を使用。
https://www.reddit.com/r/Qwen_AI/comments/1xxdpul/qwen4max_is_here_how_competitive_is_it/

［22］Reddit r/singularity. Qwen4-Max Your New Teammate Has Arrived.
2026 年 11 月 16 日投稿。閲覧日：2026 年 11 月 16 日。公開直後の反応。検索取得で確認した範囲を使用。
https://www.reddit.com/r/singularity/comments/1xxdj2b/qwen4max_your_new_teammate_has_arrived/

［23］QwenCloud. Permissions and usage policy.
公開日・更新日の明示を確認できず。閲覧日：2026 年 11 月 16 日。Token Plan の用途制限、個人版とチーム版のデータ利用条件。
https://docs.qwencloud.com/token-plan/token-plan-permissions-and-usage

［24］Alibaba Cloud. Alibaba Cloud Model Studio Data Retention Policy.
2026 年 11 月 13 日更新。閲覧日：2026 年 11 月 16 日。標準保持、ZDR の適用範囲、対象モデル、設定、例外。
https://www.alibabacloud.com/help/en/model-studio/data-retention-policy

［25］QwenCloud. Regions deployment scopes and endpoints.
公開日・更新日の明示を確認できず。閲覧日：2026 年 11 月 16 日。保存リージョンと推論実行範囲の違い、Global の範囲。
https://docs.qwencloud.com/developer-guides/getting-started/regions