Qwen4-Max 戦略的解体新書

「ハイプ」の解剖と、日本企業にとっての真のビジネス・インパクト

DOCUMENT CLASSIFICATION: EXECUTIVE BRIEFING
DATE OF ASSESSMENT: 2026.11.16
SUBJECT: QWEN4-MAX (ALIBABA CLOUD) / MODEL ID: qwen4-max
TARGET: CIO / CTO / AI STRATEGY LEADERS

Gemini Notebook

絶対性能の追求から、エージェント業務の「価格破壊」へ。Qwen4-Maxはパラダイムシフトの引き金となる。

【RELEASE DATA】

2026年11月16日正式公開。API（QwenCloud / Model Studio）限定提供。前世代（Qwen3.8-Max）と異なり、重み（Weights）は非公開のプロプライエタリ・モデルへ回帰。

【ARCHITECTURE】

総パラメータ**「2.4兆」**の巨大MoE構造。しかしトークンあたりの活性パラメータは**「710億」**に抑え込み、さらに**3,280億**のN-gram埋め込みを併用し推論効率を極大化。

【COST DESTRUCTION】

単タスクの平均費用を前世代比で**「約60%削減」**。100万トークン入力あたり**1.5ドル**（Claude Opus 5.5の約3分の1）。

【MARKET POSITIONING】

Artificial Analysis総合指数は「43点」。中国モデルトップクラス（Vals Index中国1位）だが、絶対性能ではOpus 5.5（59点）に及ばない。これは「敗北」ではなく「戦う土俵の変更」を意味する。

Gemini Notebook

「Claude Opus超え」という見出しの裏にある、性能評価の真実。

MYTH
メディア・公式主張

「最強のAI誕生・Claude Opusを撃破」

DeepSWE v1.1やHLAB-Vals（法務）でClaude Opus 5.5を上回るスコアを記録。

2.4兆パラメータの超巨大モデルによる圧倒的知能。

VS

REALITY
第三者機関の分析

「特化型エージェントとしてのコスト特化モデル」

Artificial Analysis指数では43点（同価格帯81モデル中38位）。Opus 5.5（59点）やKimi K3.1 Pro（46点）には届かず。

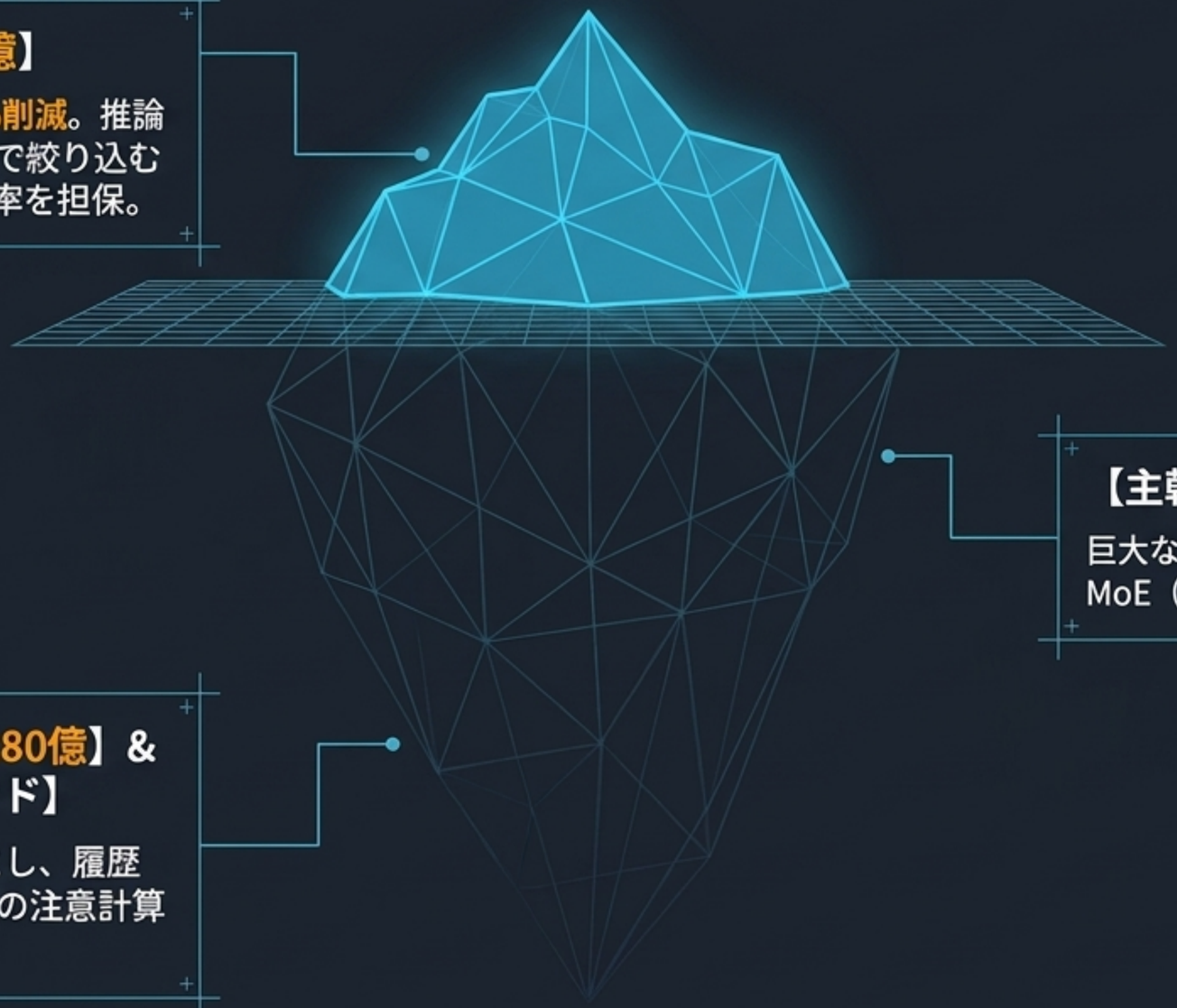
16の比較パネル中、QwenがOpusを上回るのは6項目のみ。ターミナル操作やプログラミング（Terminal-Bench 4.0等）では明確な劣後。

Gemini Notebook

なぜ2.4兆パラメータの巨大モデルが、前世代より高速かつ低コストで動くのか？

【活性パラメータ：710億】

前世代（約950億）から**約25%削減**。推論時に実際に動く部分を極限まで絞り込むことで、生成速度とコスト効率を担保。



【主幹パラメータ：2.4兆】

巨大な知識ベースとして背後に控えるMoE（Mixture of Experts）構造。

【N-gram埋め込み：3,280億】&【GDN + QSAハイブリッド】

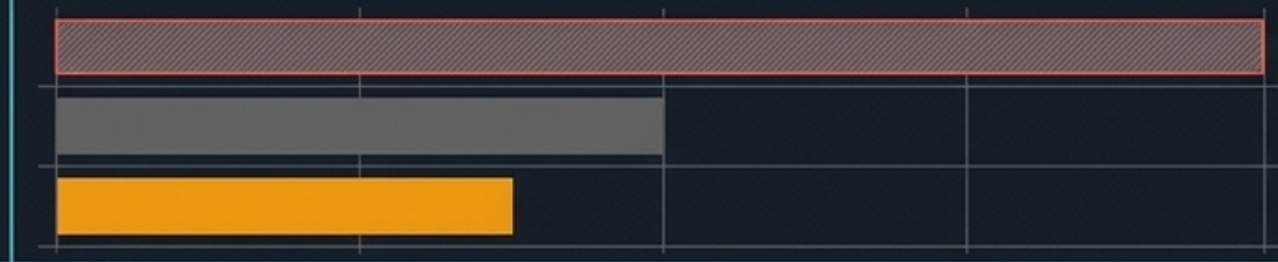
4層中3層をGated DeltaNetとし、履歴を固定長に圧縮。長い系列での注意計算コストを**劇的に抑制**。

Gemini Notebook

エージェント業務を前提とした、桁違いのトークン・エコノミクス。

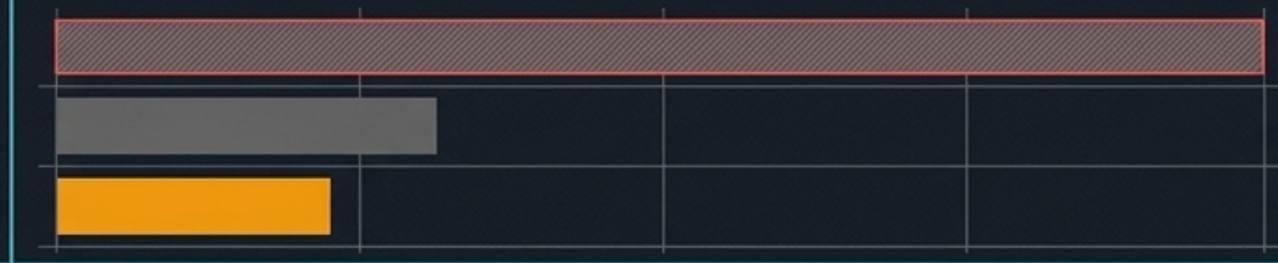
100万トークンあたりの国際API価格比較

【入力単価】



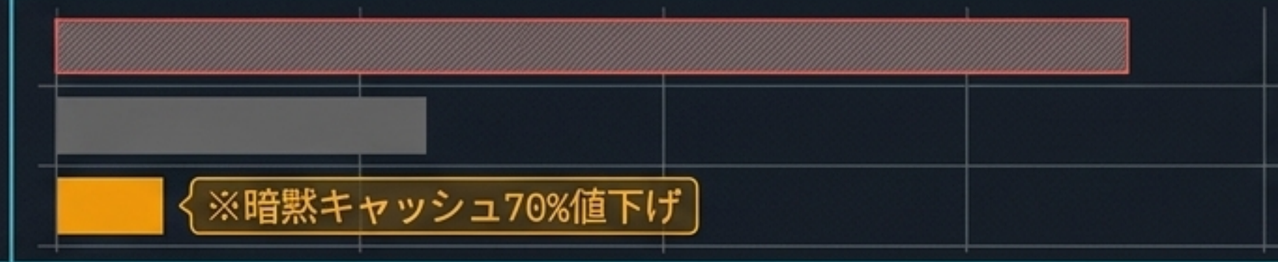
Claude Opus 5.5: $4.00
Qwen3.8-Max: $2.00
Qwen4-Max: $1.50

【出力単価】



Claude Opus 5.5: $20.00
Qwen3.8-Max: $6.00
Qwen4-Max: $4.50

【キャッシュヒット】



Claude Opus 5.5: $0.20
Qwen3.8-Max: $0.25
Qwen4-Max: $0.075

インサイト・ボックス

トークン単価の低下に加え、アーキテクチャ改良により「トークン消費量自体」も減少。Terminal-Bench 4.0ではl試行あたりの平均消費トークンが289Kから151Kへ半減。 これが「単タスク費用60%減」の正体。

Gemini Notebook

万能の天才ではなく、特定業務に特化した「鋭利なツール」。

強み・推奨領域（Strong / Recommended）

【法務エージェント】
HLAB-Vals 1.1: 21.3（Opus 9.2） -
大量ドキュメントの読み込みと特定ルールの適用。

【生物・医療・先端科学】
BioMysteryBench: 85.0（Opus 81.9）

【文書・ツール利用】
100万トークンの巨大コンテキストを活かした抽出・要約作業。

弱み・非推奨領域（Weak / Not Recommended）

【長時間のターミナル操作】
Terminal-Bench 4.0: 48.48%（Opus 65.15%）

【自律的コーディング】
ProgramBench: 31.5（Opus 65.0）

【セキュアコード生成】
完全な自律開発を任せるには、依然として米国の旗艦モデルに軍配。

Gemini Notebook

パラダイムシフト — 「絶対知能の王者」から「エージェント・ワークフローの最適解」へ。



Qwen4-Maxの設計思想は、ベンチマーク上の1位を獲ることではありません。 活性パラメータの意図的な削減とN-gramの採用は、**「推論コスト」**と**「消費メモリ」**を極限まで抑えるためのGPU制約の産物であり、同時に長時間の反復タスク（エージェント業務）を商業的に成立させるための戦略的ピボットです。

戦略的意義：日本企業は、このモデルを**「単発の賢いチャットボット」**として評価すべきではありません。数十回のツール呼び出しや思考プロセスを繰り返す**「自動化ワークフローのエンジン」**として、**ROIを測る**べきです。

Gemini Notebook

日本企業における導入の壁：コンプライアンスとデータ・レジデンシー。



RED（クリティカルなリスク）：【データ処理リージョン問題】

API対応リージョンに「東京」が含まれるが、デプロイ範囲は「Global」と定義されている。東京で受けたリクエストが国外（中国本土等）で推論・処理される可能性が高く、 機密データの投入には厳格な契約確認が必須。



YELLOW（運用上の制約）：【オンプレミス・ファインチューニング不可】

前世代のMaxと異なり重みは非公開（クローズド）。自社環境へのデプロイや学習データの追加は全リージョンで非対応。個人向けToken Planはシンガポール限定。



GREEN（セキュアな設計思想）：【学習時の安全性強化】

コードインジェクション、メモリ安全性、CBRN（化学・生物・放射性物質・核）向けの悪用対策データフィルタリングが強化されている。

Gemini Notebook

特定ユースケース診断：機密業務（知財・法務）への適用限界。

ユースケース	強み	弱み	最終判定
大量の引用文献・明細書の要約	100万トークンの巨大コンテキスト。	出願前の機密情報の場合、APIの「Global」ルーティングによる情報流出リスク。	FAIL （情報セキュリティ観点で本番投入は保留）
法的なリサーチと論理的判断	HLAB-Valsでの高スコア。	Legal Research Benchでは41.35%（Opusは50.48%）。高度な推論を完全に任せる水準には未到達。	CAUTION （人間の専門家を補助する「下書き・抽出ツール」としての利用に限定）

Gemini Notebook

次の一手 — 戦術的ROIの検証とオープンモデルへの備え。

01

本番移行前の「タスク単価ROI」検証

トークン単価の安さだけでなく、「手戻り」を含めた最終的なタスク完了コストをClaude Opus 5.5等の社内標準モデルと比較する。

プレビュー版（11/25終了）から正式版APIへ即時切り替える。

02

エージェント型コーディング業務の切り離し

リポジトリ規模の自律開発や長時間のターミナル操作には適さない。
コーディングの主幹エンジンは米国の旗艦モデルを維持し、Qwen4-Maxはドキュメント処理や補完的な自動化ワークフローに配置する。

03

オープンウェイト小型版（27B / Flash）の待機

知財や機密データを扱う企業は、現行のMax APIへの機密データ投入を避けよ。
今後予告されている小・中規模モデルの重み公開（オープンソース化）を待ち、自社オンプレミス環境でのローカル検証を準備する。

Gemini Notebook