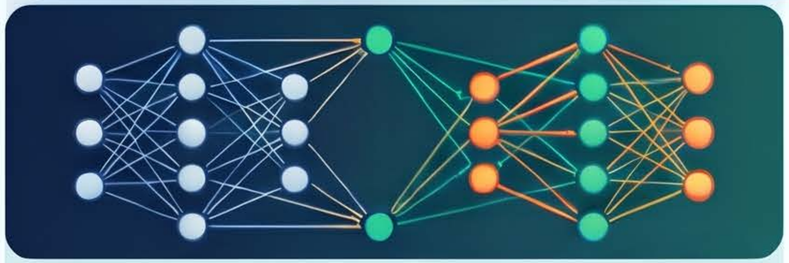
Qwen4-Max 徹底解剖：性能・コストと日本企業の導入留意点



活性710億の効率的MoE構造

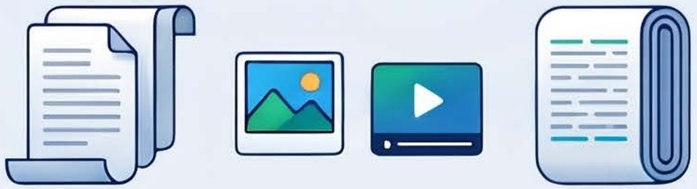


主幹2.4兆パラメータのうちトークン活性を710億に抑え高速化。



主幹22.4兆パラメータのトークン活性を71億に抑え高速化。

100万トークンの長文脈対応



テキスト・画像・動画入力に対応し、長大ドキュメントの解析に最適。

タスクコストを約60%削減



前世代モデル



API価格は前世代比25%引き下げ、100万入力あたり1.5ドルを実現。

Qwen4-Max（国際版）	Qwen3.8-Max（前世代）	Claude Opus 5.5
入力価格：$1.50	入力価格：$2.00	入力価格：$4.00
出力価格：$4.50	出力価格：$8.00	出力価格：$20.00
キャッシュ価格：$0.075	キャッシュ価格：$0.250	キャッシュ価格：$0.200

性能評価と日本企業の導入チェックポイント

専門領域に強み・自律開発に弱み



法務・生物分野：
圧倒的な強み

長時間のエージェント型
コーディング：苦手

法務や生物分野で圧倒する一方、
長時間のエージェント型コーディングは苦手。

重み非公開のクローズド提供



オンプレミス導入やファインチューニングは
全リージョンで非対応。

東京リージョンでもデータ所在確認が必須



クロードデプロイ



デプロイ範囲：Global 表記

デプロイ範囲が「Global」表記のため
機密データ入力前の契約確認が不可欠。

Gemini Notebook