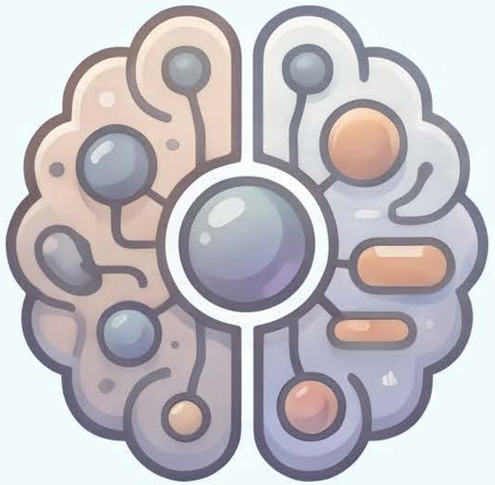
Qwen4-Maxの性能・コスト評価と実務導入の留意点

アーキテクチャ進化とコスト効率



Qwen3.8-Max

計算効率の
着実な向上



Qwen4-Max

基幹2.4兆パラメータに対し活性化規模を710億へ削減し処理を効率化

API利用料金の値下げ



総合性能と費用の改善



第三者指数が34から43へ上昇

評価タスクあたりの費用



実務応用力とガバナンス課題

法務・開発分野での強み



法務ベンチマークの
個別項目合格率91.2%



Web開発評価で
上位を獲得

単価と総費用の乖離



思考トークンの伸長により単価が安くても総費用が増える場合があり注意が必要

データ保持と推論地域の確認



東京リージョン指定時も推論実行ノードが国外に及ぶ可能性があり確認が必須

Gemini Notebook