Qwen4-Max-Preview 徹底解説：仕様・性能・実装ガイド

2026年11月に15日間限定の公開ストレステストとして提供されたQwen4-Max-Previewは、100万トークンの入力長と前世代の3倍となる39.3万トークンの共有出力枠を備えた超大型推論モデル。高い自律コーディング能力と長文脈検索精度を誇る一方、Token Plan限定の利用規約や出力切断を招くAPI仕様変更などの留意点が存在する。

基本仕様とアーキテクチャの革新



出力上限

思考と回答で枠を動的共有。





3対1の複合機構

8.6倍の高速文脈処理を実現。





510億メモリ層

FLOPsゼロで知識容量を拡張。



世代間比較

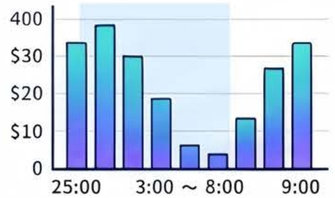
比較項目	Qwen4-Max-Preview	Qwen3.8-Max	Qwen3.8-Flash-Next
最大出力枠	39.3万トークン	13.1万トークン	思考26.2万+困箋13.1万
思考モード	Hybrid (xhigh既定)	Hybrid (既定ON)	Hybrid (xhigh既定)
提供形態	Token Plan限定	従重課金API / OSS	オープンウェイト

利用経済性と実装上の注意点



夜間92%オフ

日本時間23時～翌9時が格安。





帰納ルール発見の強み

接継な価格検証で覗合を凌駕。



max_tokensの変更

思考と回答の合算値指定に注意。

Gemini Notebook