

Alibaba Qwen3.8-Max：2.4兆パラメータが拓く次世代AIの性能と実態

基本仕様と圧倒的な性能



Sparse MoE構造

推論時の有効パラメータは95Bで、効率と性能を両立したネイティブマルチモーダルモデル。

100万トークンの超長文コンテキスト
最大13万トークンの出力長をサポートする。



100万トークンの超長文コンテキスト

LMarenaでVision部門2位、PaperBench(93.0)などで競合他社を凌駕する。

実務導入の可能性と潜在的リスク



Qwen3.8-Max



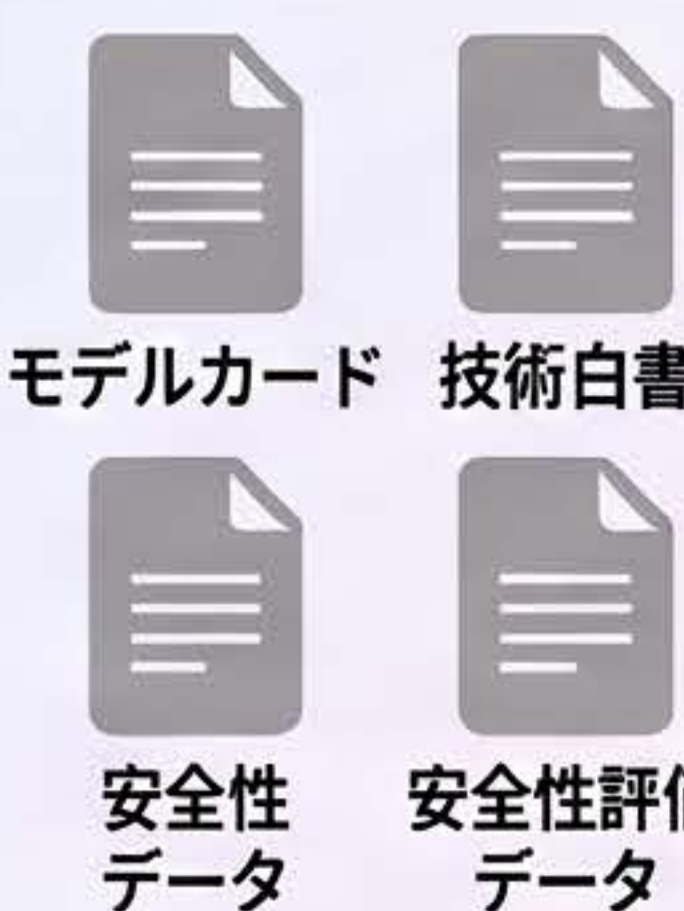
Claude Opus 4.6

競合を圧倒する破壊的な低価格

入力100万トークンあたり2ドル。
Claude Opus 4.6(5ドル)の半分以下で提供。

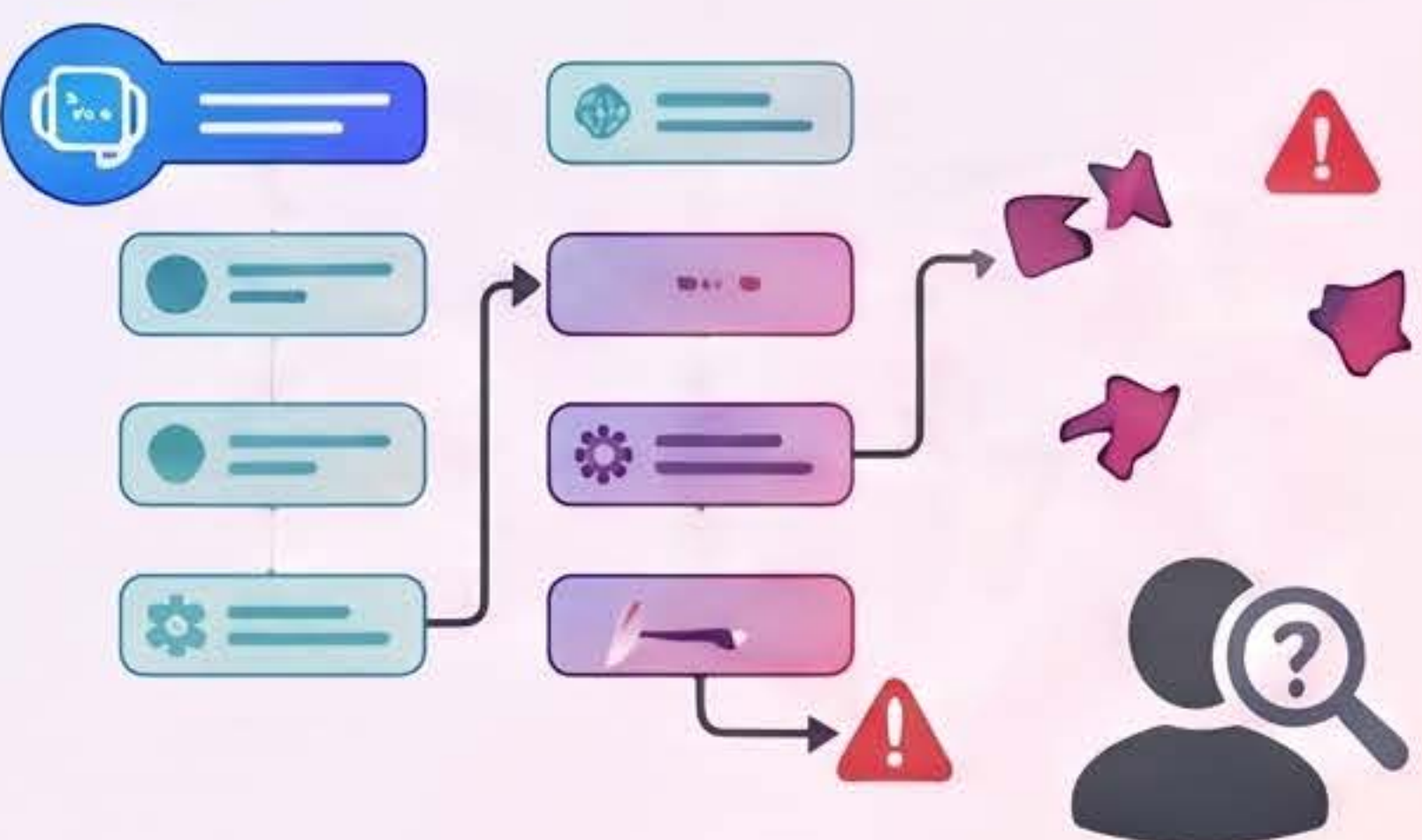
世界トップクラスのベンチマーク順位

主要ベンチマークにおける競合モデルとの性能比較 (Alibaba報告値)



技術的透明性と安全性の不足

モデルカードや技術白書、安全性評価データが未公開であり、全面移行には慎重な判断が必要。



長時間エージェントでの信頼性課題

長期セッションでツール呼び出しの構造が崩れる報告があり、人間による確認が必須。