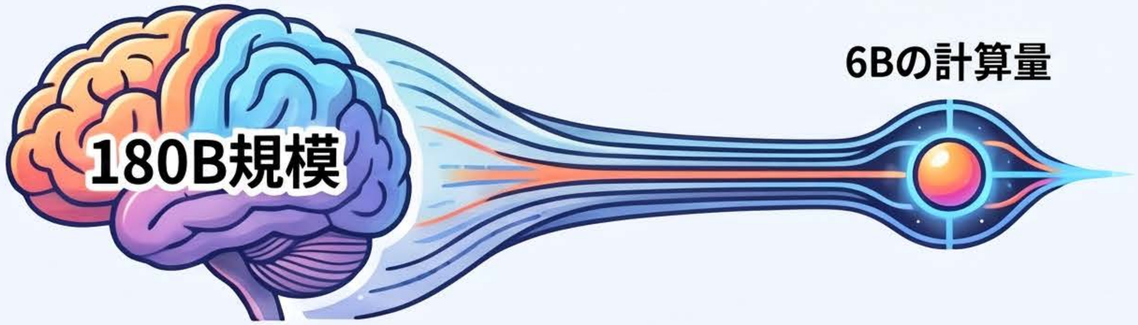
Qwen3.8-Flash-Next：次世代AIアーキテクチャの全貌と実務ガイド

技術的革新：Qwen4への道



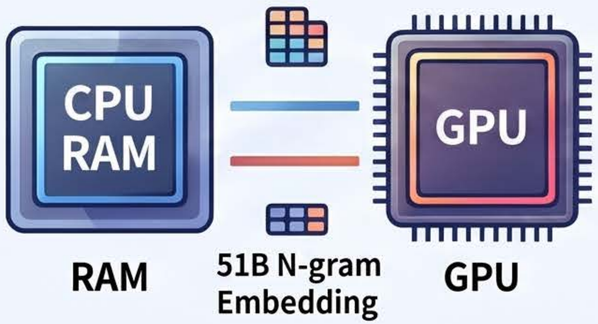
180B相当
(総パラメータ数)

125BのMoEモデルのうち1トークン当たり約6Bのみを活性化し、推論を高速化。

長文処理を加速する GDN+QSA

micro-block単位で重要領域を選択し、
1Mトークンの長文処理を最大8.6倍高速化。

6B parameters
(アクティブ計算量)

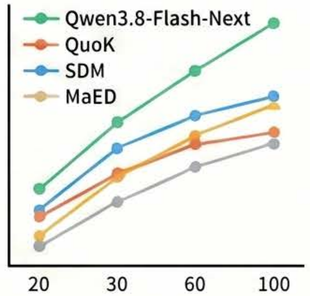


CPU RAMを活用する N-gram Embedding

51Bのパラメータをホスト RAMへオフロードし、GPU負荷を抑えつつ知識容量を拡大。

最大1,000,000 tokens
(コンテキスト長)

実務導入の検討事項とリスク



コーディング・推論で高い性能を発揮

SWE-bench等で既存モデルを凌駕するが、一部のベンチマークはベンダ報告値に留まる。

Community License 1.0の制約

Apache 2.0ではなく、MaaSやAIアシスタント事業には別途ライセンスが必要。

運用リソースの現実

計算量は少ないが、重み一式の保持に約180GB〜360GBのメモリ空間を必要とする。

Gemini Notebook