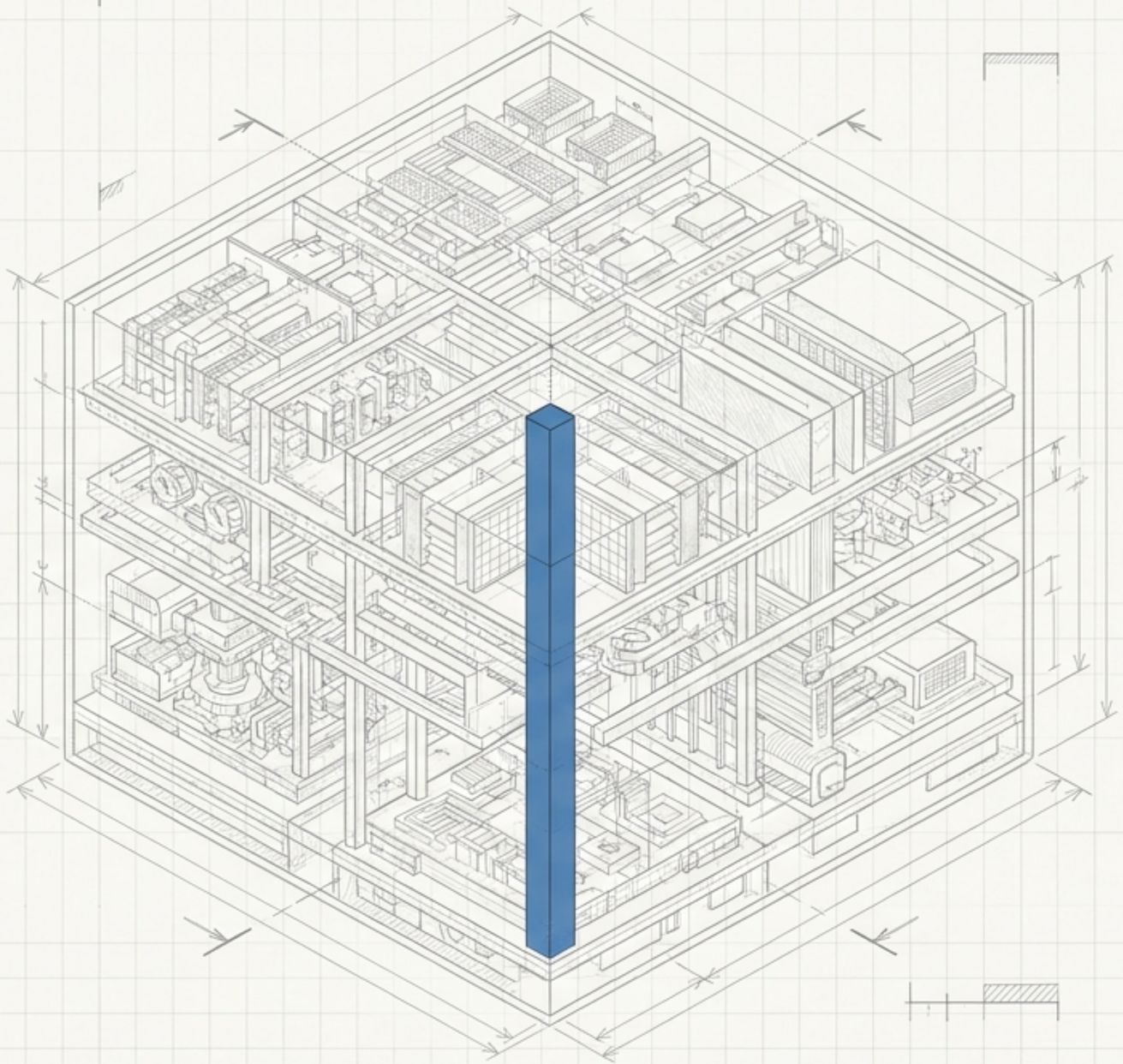
次世代アーキテクチャ「Qwen4」への青写真

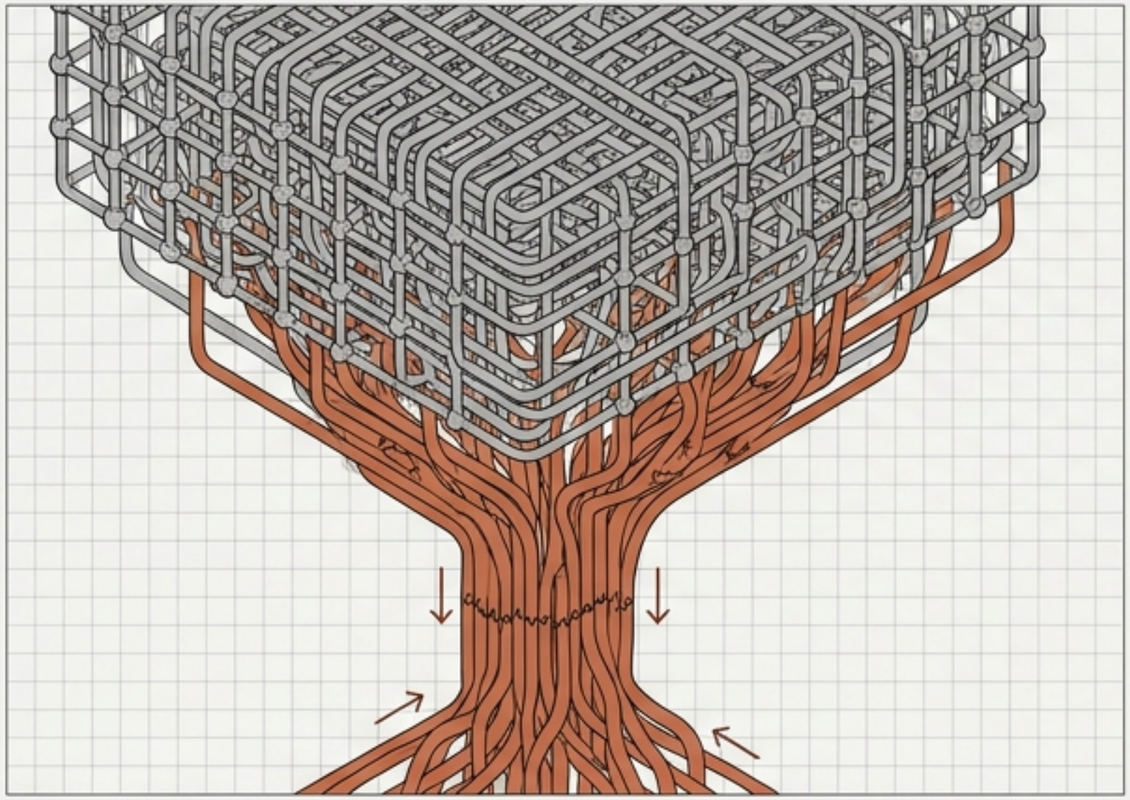
Qwen3.8-Flash-Next：演算と記憶を完全分離するUltra-Sparse MoEの衝撃



August 2026 | Technical Briefing

巨大化する知識と推論コストのジレンマ

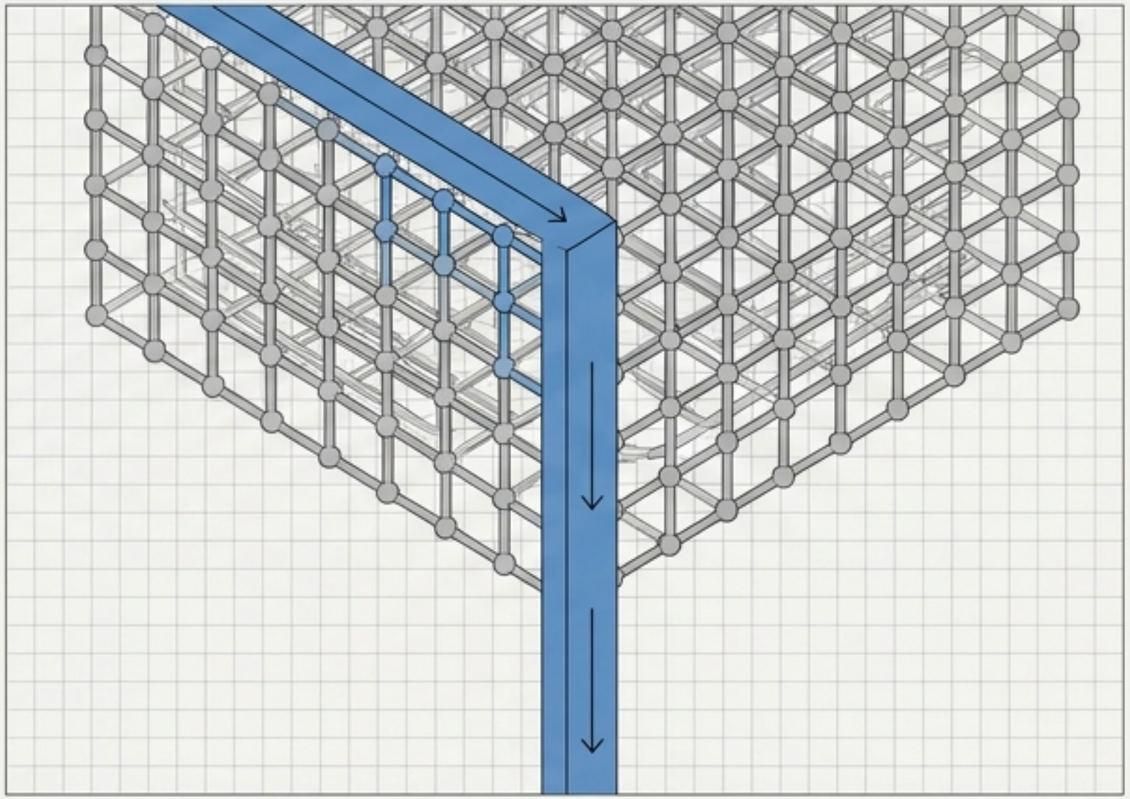
従来のDenseモデル（ボトルネック）



業界の壁

パラメータ拡大に伴う、計算資源とメモリ帯域の深刻な枯渇

Qwen3.8-Flash-Next（根本的解決）



Qwenのブレイクスルー

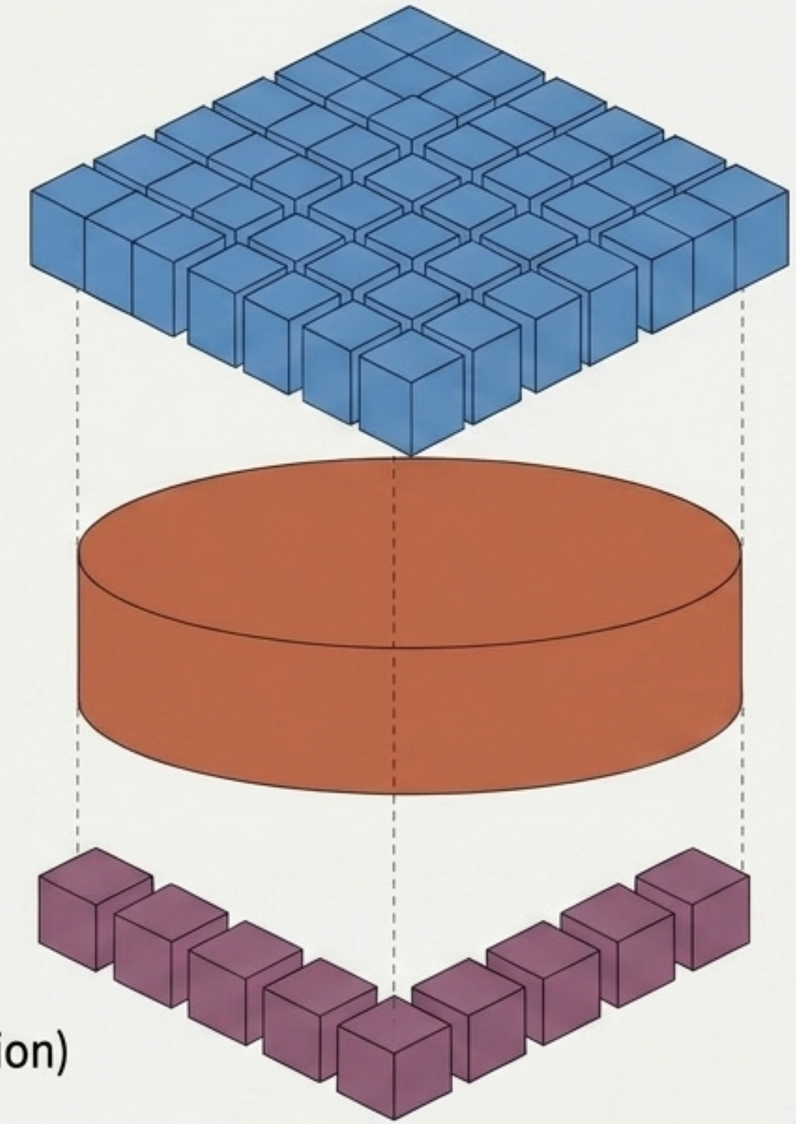
総パラメータ125Bに対し、推論発火をわずか「6B」に極小化。圧倒的知識量と推論コストの矛盾を両立。

アーキテクチャの三権分立：演算と記憶の構造的分離

演算領域
(Ultra-Sparse MoE)
125B Backbone (Active 6B)

長期知識領域
(N-gram Embedding)
51B Table (Offloaded)

長大文脈検索領域
(GDN+QSA)
4B MTP (Context Compression)



モノリシック設計からの脱却

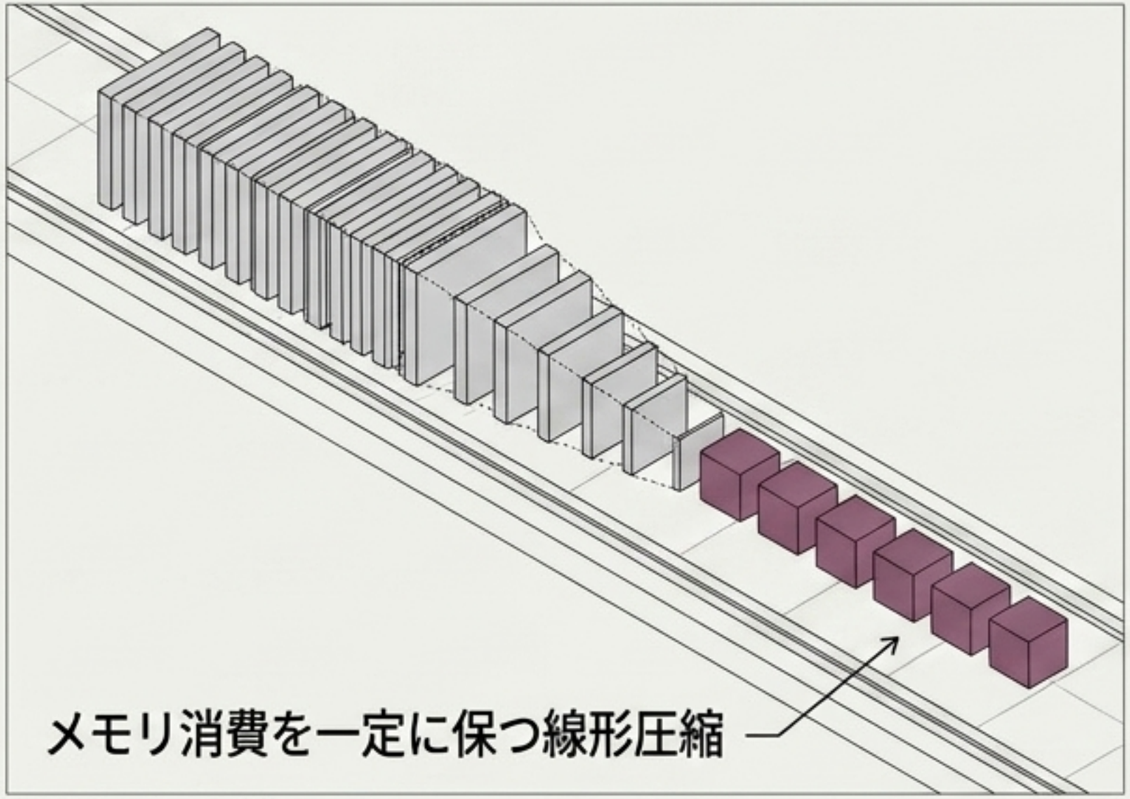
テキスト・画像・動画の高度な理解をネイティブで維持しながら、モデルの物理的・論理的な役割を完全に分離。

総パラメータ1250億（125B）に対し、稼働率を約5%（6B）に抑え込むことで、次世代の極小推論コストを実現。

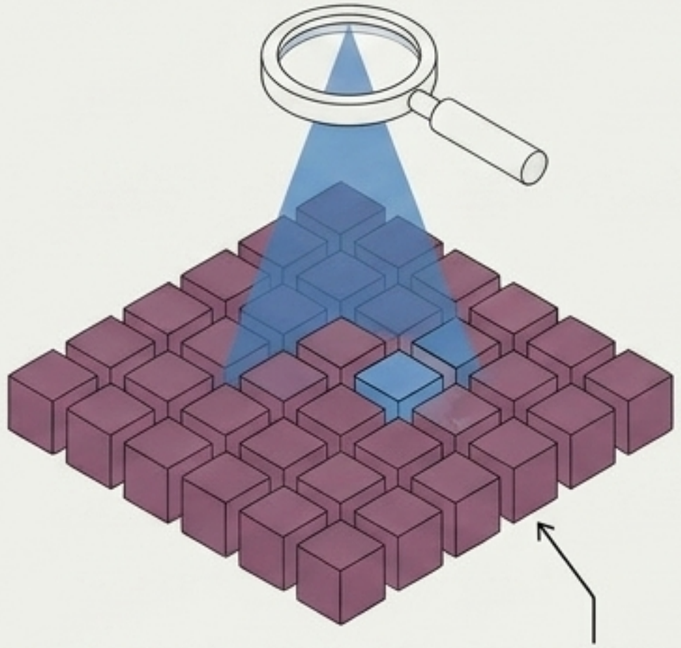
圧倒的な知識量と推論コストの矛盾を根本的に解消。

ハイブリッド・アテンション：長大文脈の圧縮と高精度検索

GDN (Gated DeltaNet) - 36 Layers



QSA (Qwen Sparse Attention) - 12 Layers

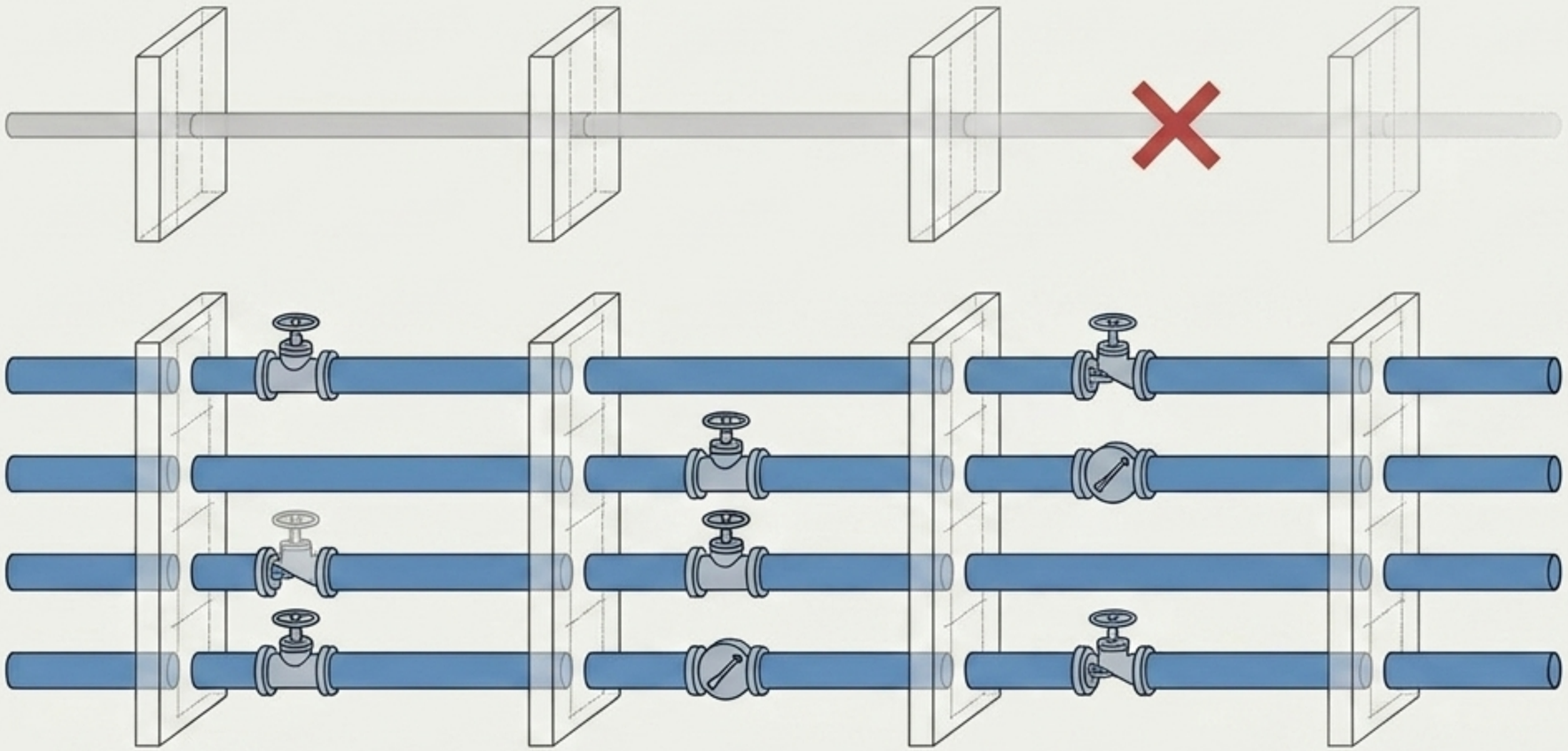


Prefill スループット向上

8.6倍

プレフィックスキャッシュ命中率90%・
100万トークン入力環境時 (Qwen3.7-Plus比)

Gated Residual：深層への「Hyper-Connections」



従来の単一バイパス
(深層での勾配消失・情報希釈)

Hyper-Connections
(4並列の動的ルート)

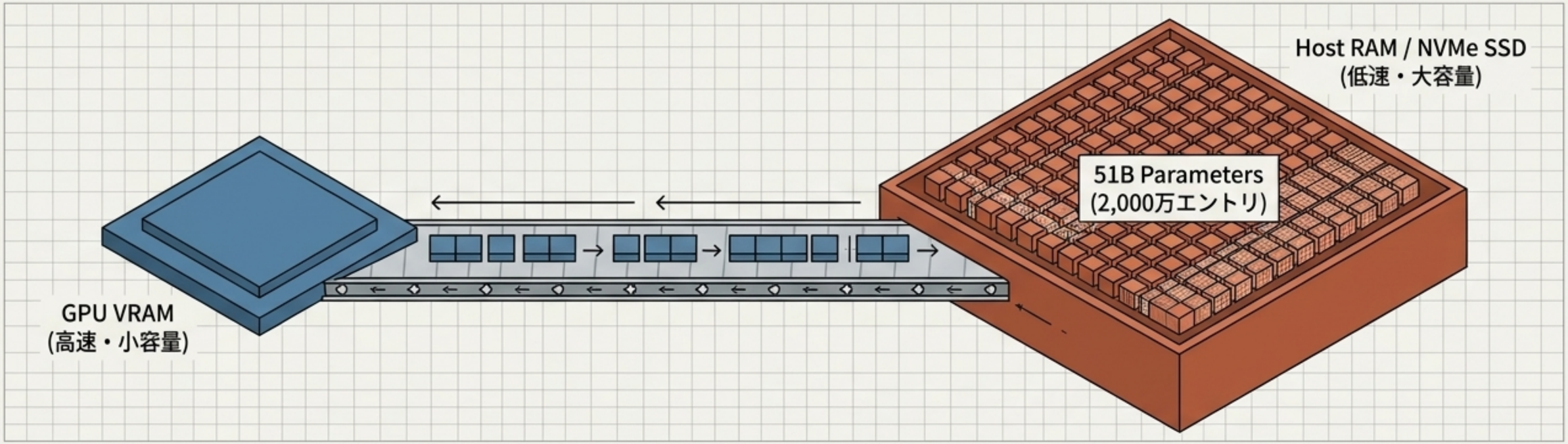
動的ゲート制御

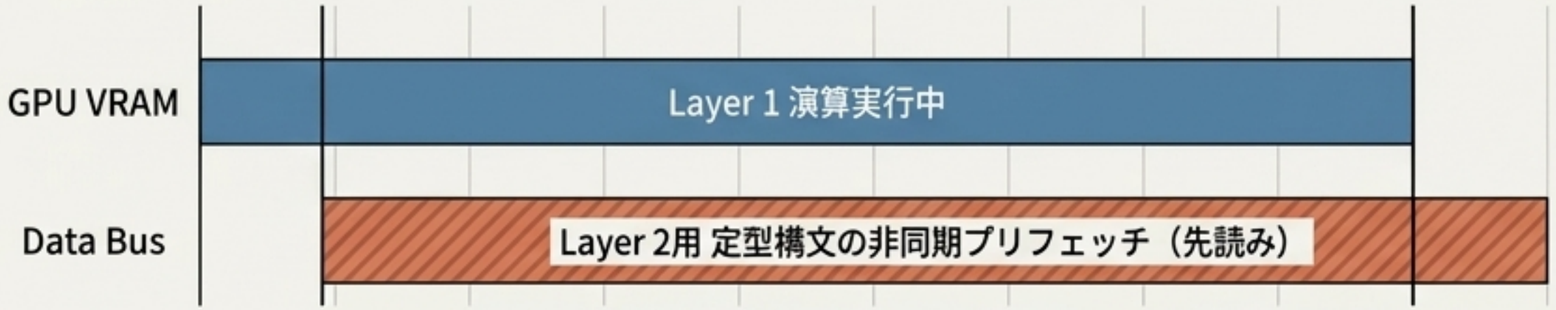
リード/ライトゲートにより、入力データに依存して情報の読み書きを要素・ブランチ単位で動的に制御。

情報伝達効率の飛躍

初期層で抽出された局所的な文脈情報が深層で希釈されるのを防ぎ、推論オーバーヘッドを最小化しながら数値を安定化。

N-gram機構：静的知識のSSD/RAMオフロード

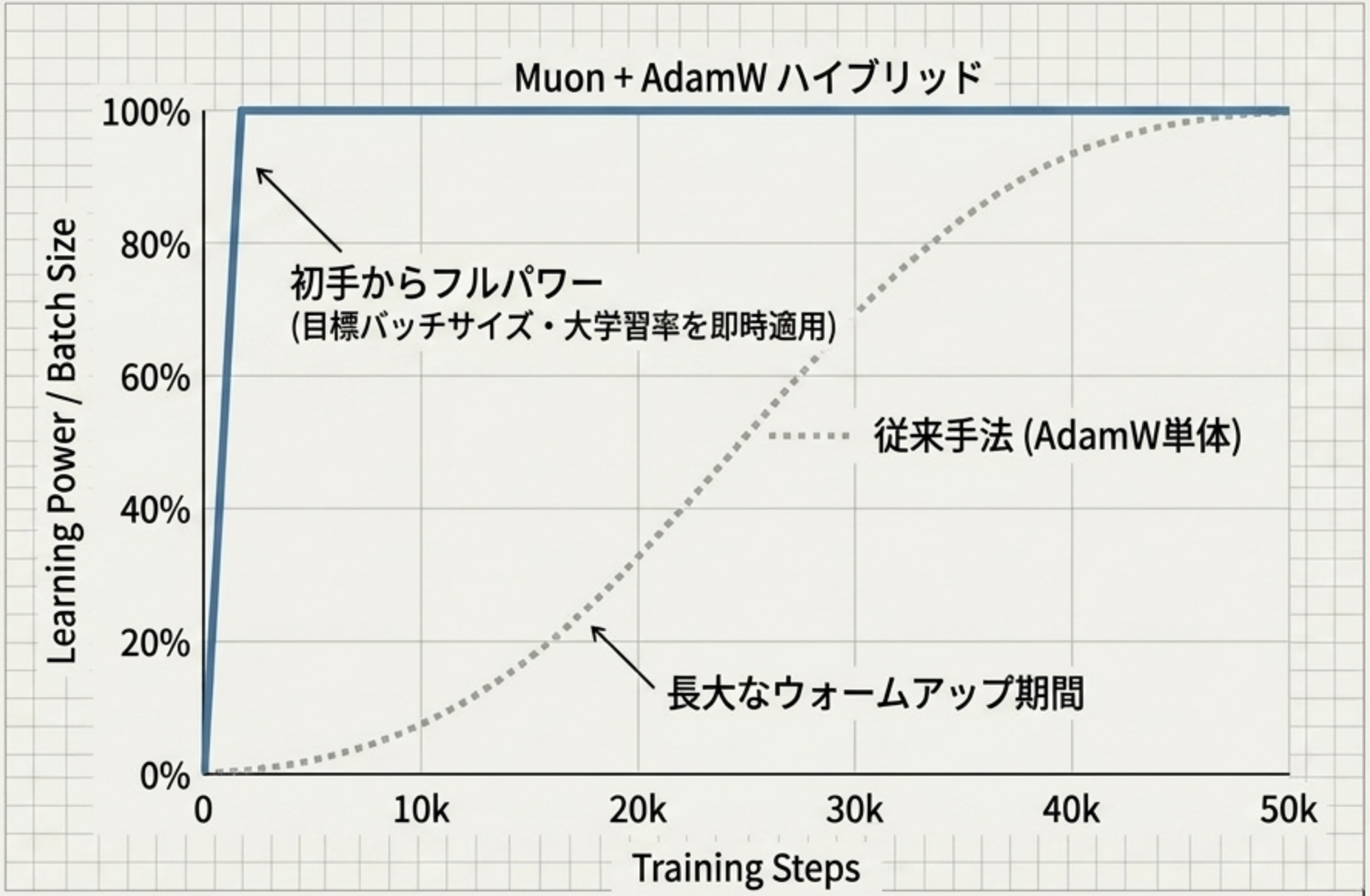




最大のゲームチェンジャー

高価なGPUメモリからの脱却。頻出熟語やプログラミング定型構文の静的知識を外部化し、VRAM消費を劇的に削減しながらレイテンシ劣化を防止。

Muonオプティマイザによる学習コスト「9分の1」の実現

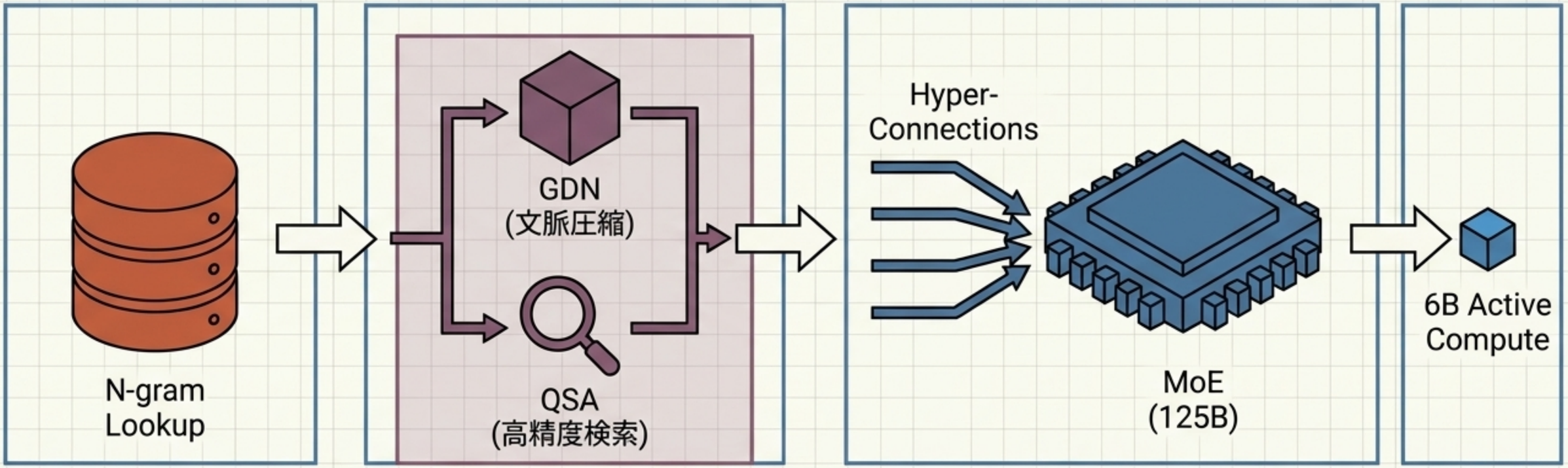


事前学習計算コスト

約 11%

前世代 (Qwen3.7-Plus) と同等能力の獲得に必要な計算資源を、1/9に抑制。行列更新の直交化制約を活用した圧倒的な学習効率。

Synthesis: 次世代推論エンジンのフォワードパス



ホストSSDから直前トークン文脈を非同期取得

ハイブリッド・アテンションによる記憶維持

125B中、6BのExpertのみが発火し推論実行

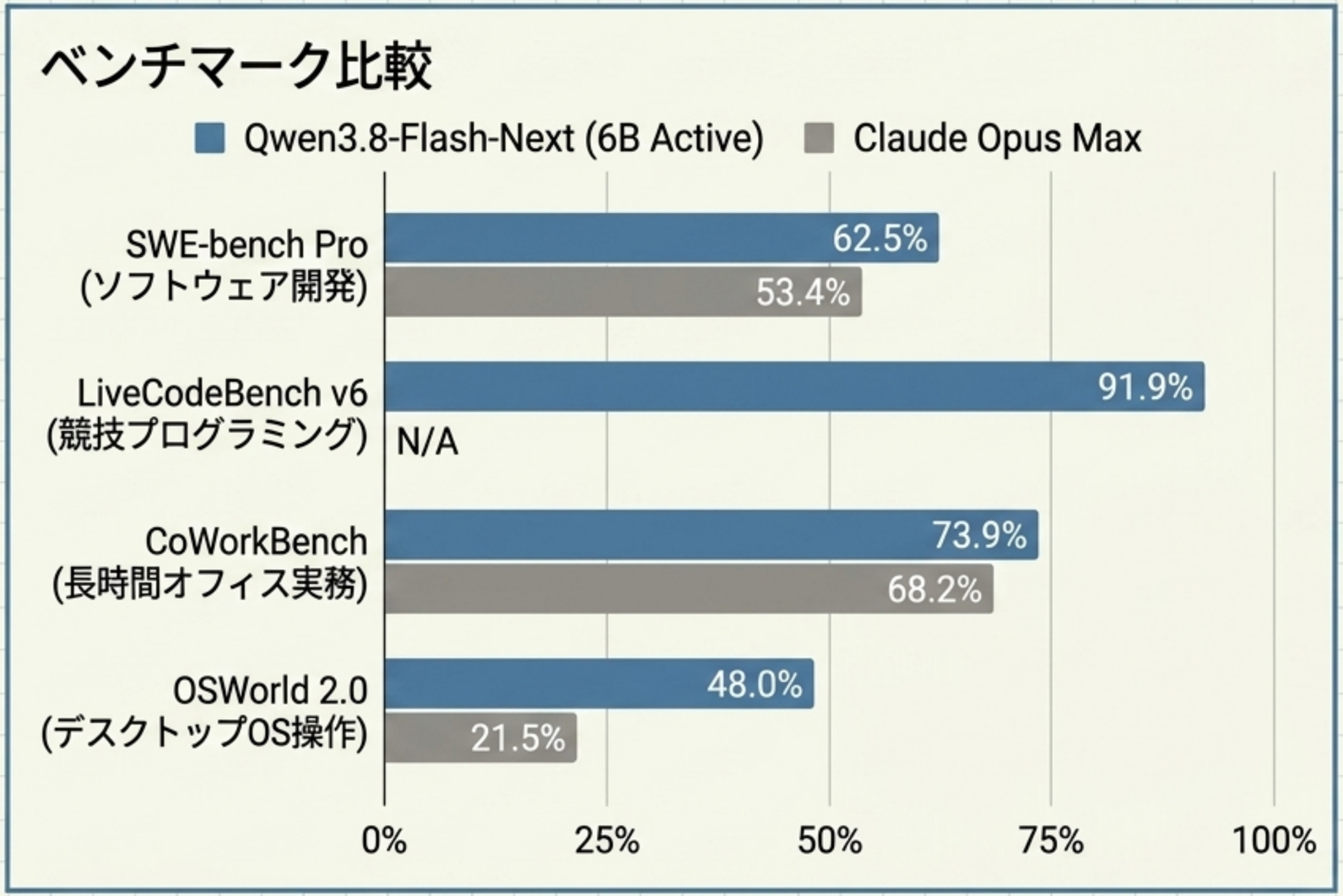
記憶は外部ストレージから引き出され、長文脈は圧縮され、極小の演算リソースのみが駆動する。これが「演算と記憶の完全分離」の全貌。

推論環境の拡張：データセンターからエッジPCまで

フォーマット	必要メモリ容量	主なターゲット環境	特徴
公式 BF16	約355〜360 GB	データセンター (GB300 NVL72, 8xH100)	スループット16,000 tokens/sec超の実証
公式 FP8	約185 GB	ワークステーション (4xRTX PRO 6000)	エンタープライズの スケールアウト検証
Unsloth UD-Q4_K_XL	約111.3 GB	ハイエンドPC (Apple Silicon M-Max/Ultra等)	128GB統合メモリ機での ローカル推論
Unsloth UD-IQ1_S	約72.5 GB	ミドルエンドPC	96GB統合メモリまたは RAMオフロード構成

Day Zero 統合完了: SGLang | vLLM | llama.cpp

フロンティアモデルとの真っ向勝負

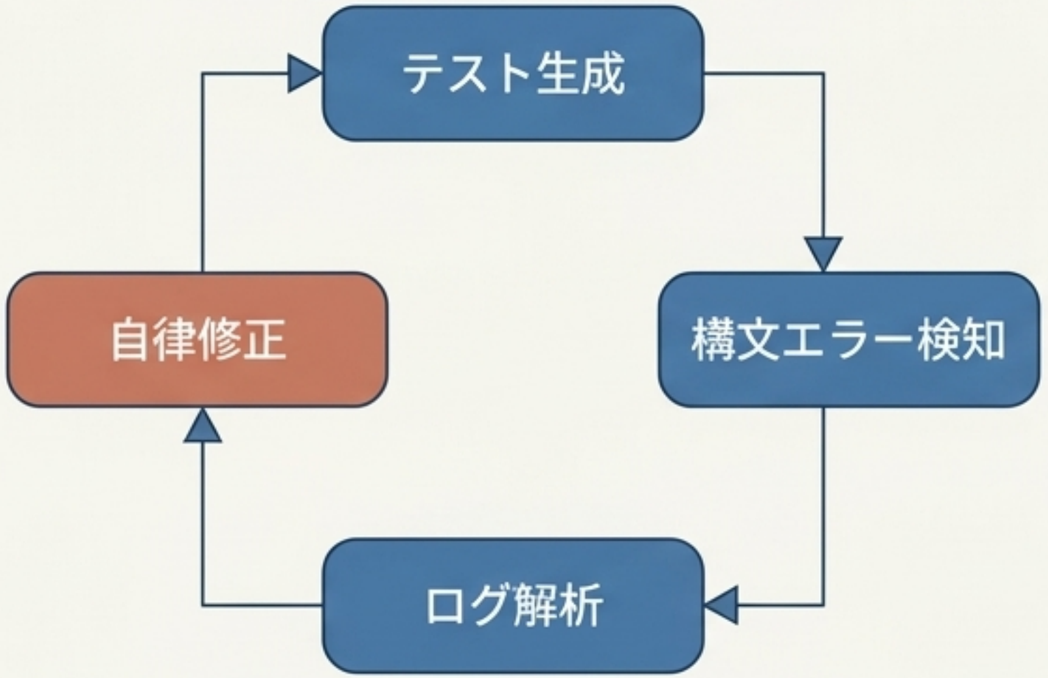


自律エージェント能力の証明

単発の知識検索ではなく、コーディングや長時間のツール実行など、実務遂行力において最上位水準を記録。6Bという極小の演算リソースで巨大なフロンティアモデルを凌駕。

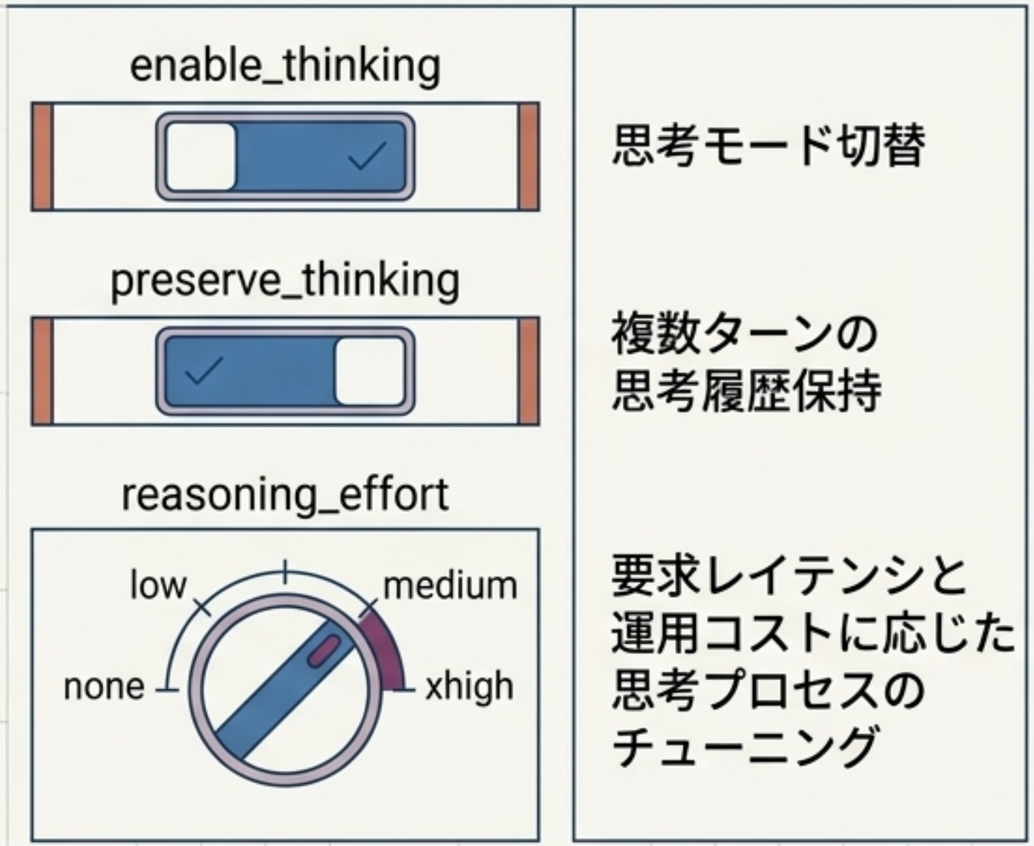
粘り強い実務遂行力と「推論深度」の動的制御

粘り強いエージェントループ

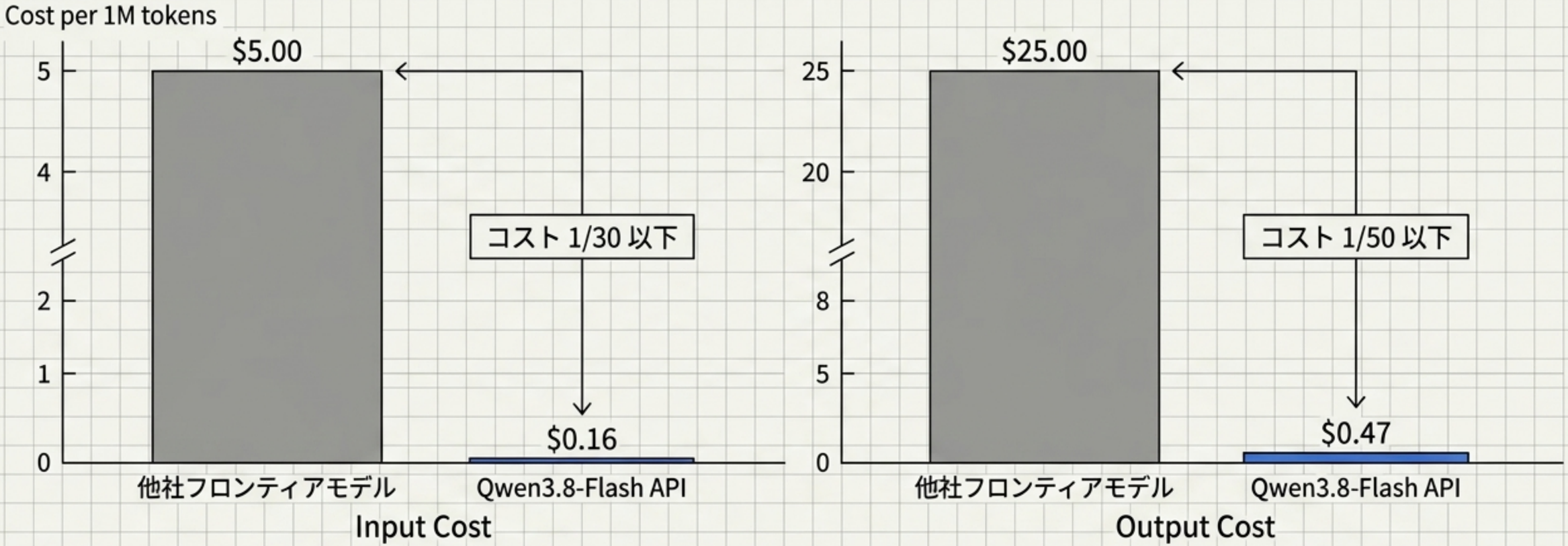


複雑なタスクにおいて、GDN/QSAがコンテキスト破綻を防ぎ、自律検証ループを完遂。

推論制御UIパネル



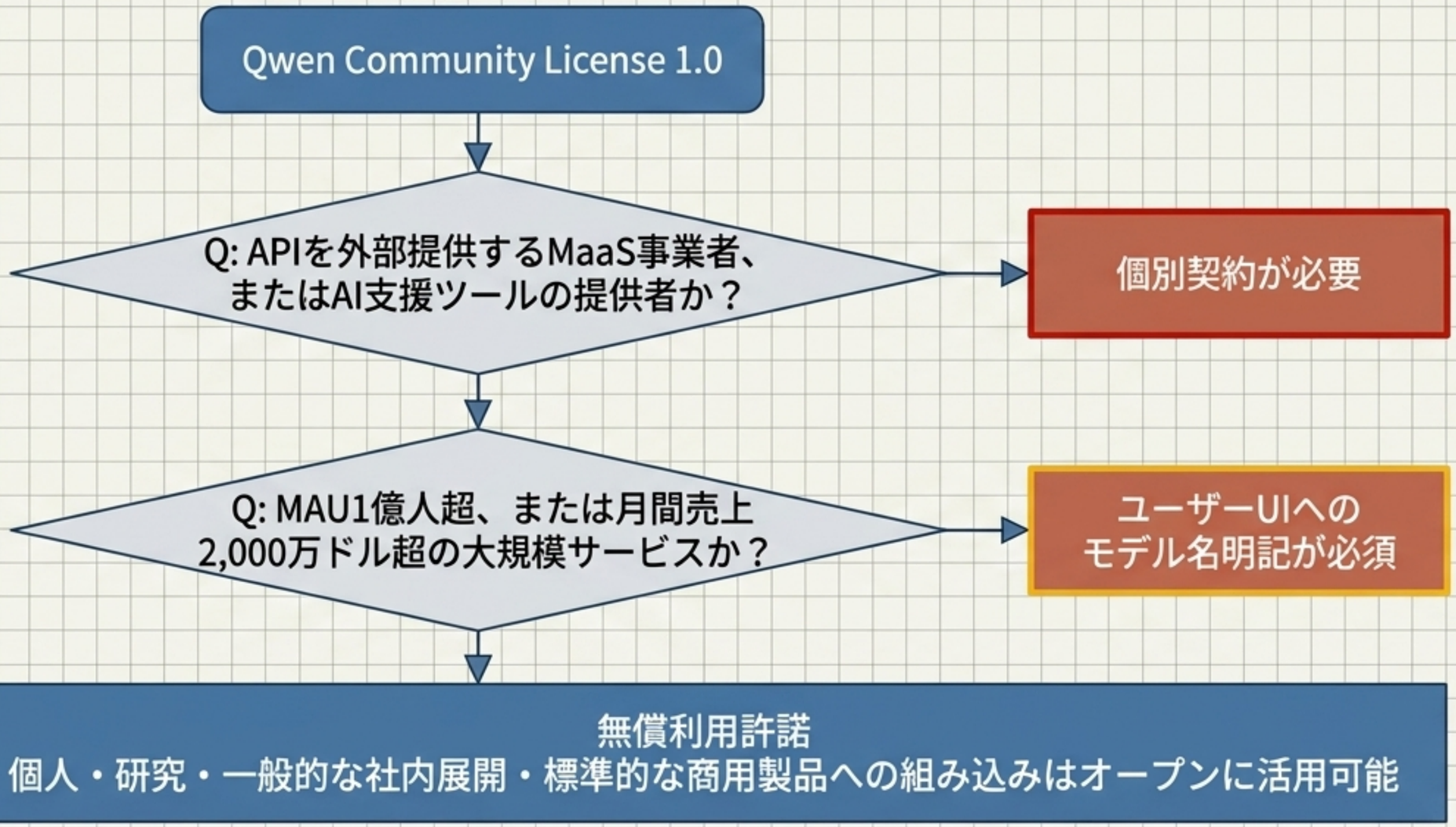
エージェント実用化を加速する破壊的なAPIエコノミクス



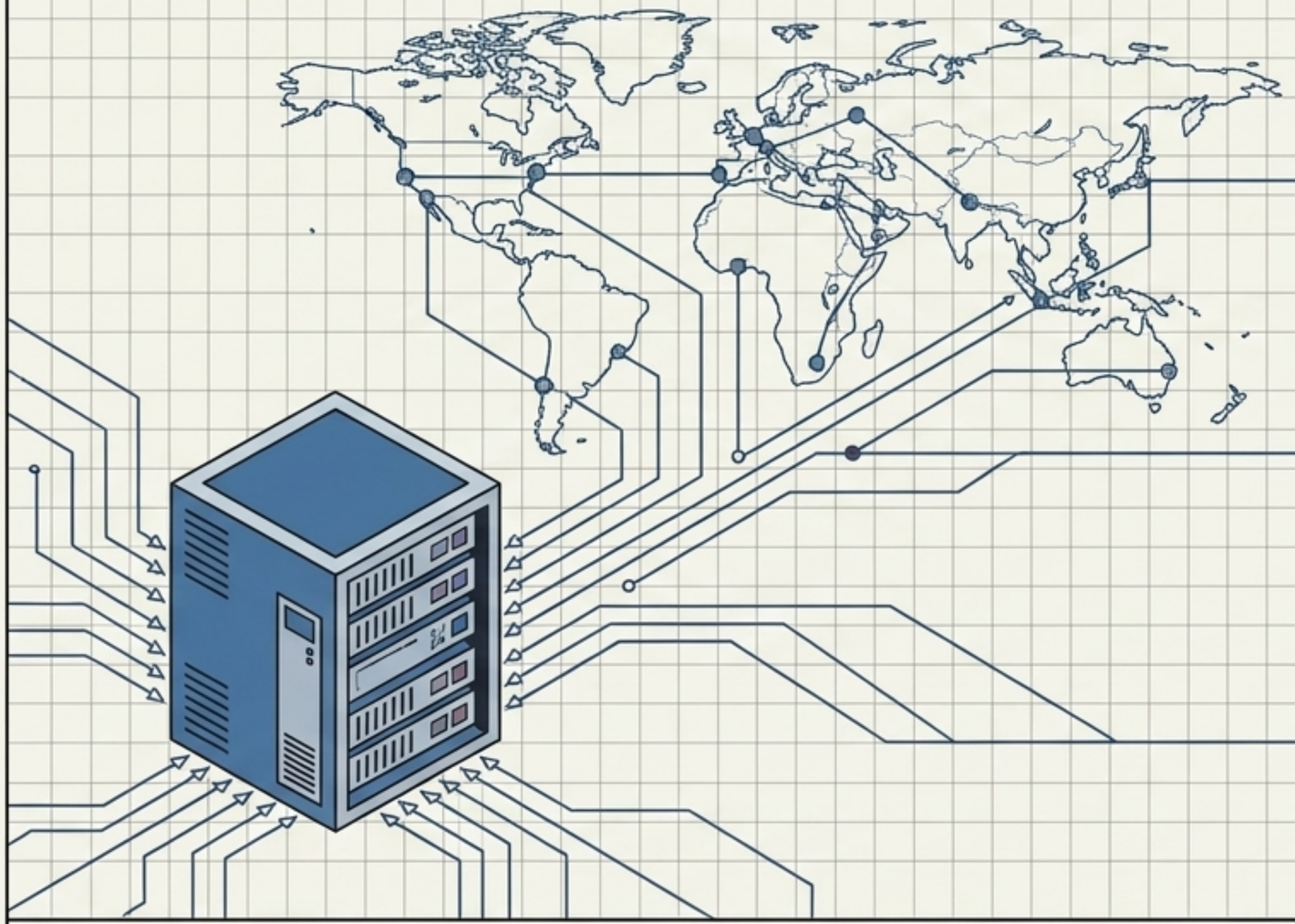
現実的な予算内での運用

入力側で30倍以上、出力側で50倍以上の圧倒的なコスト優位性。100万トークンの長文脈ウィンドウを活用した大規模バッチ処理や、高頻度のエージェント自律実行が遂にビジネスユースで実現可能に。

商用エコシステム展開のためのライセンス境界線



The Horizon: Qwen4がもたらす生成AIの構造転換



エッジへの巨大知識回帰

計算と記憶の分離により、限られたハードウェア環境でもフロンティア級の巨大AI運用が現実のものに。

真のエージェント時代の幕開け

長文脈の超低コスト維持（GDN+QSA）が、自律型AIの「粘り強い思考ループ」を経済的に支える。

来るべき「Qwen4」への確信

本技術プレビューでの成功を基盤に、ハイブリッド・疎結合アーキテクチャがさらに巨大なスケールで展開される未来。

超低コストAPIとオープンウェイトが両輪となり、生成AIの次なるパラダイムを牽引する。