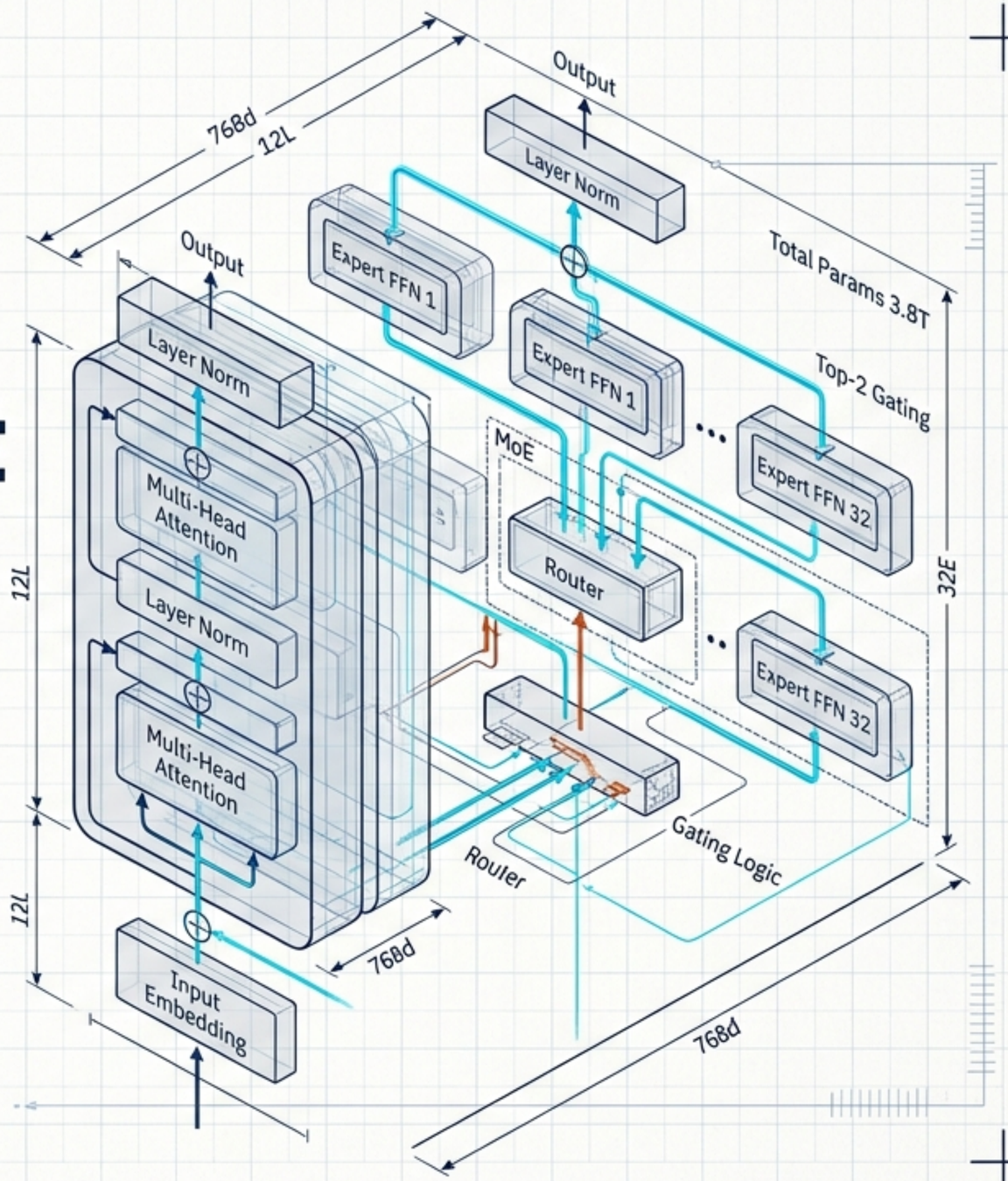
Qwen3.8-Flash-Next 徹底解剖

Qwen4に向けた次世代アーキテクチャの全貌と、自社導入のための技術的・ビジネス的リアリティ

Strategic Architecture & Deployment Analysis Report



エグゼクティブサマリ：「180Bの重さ」と「6Bの軽さ」のパラドックス

The Identity (位置づけ)

Qwen4で採用予定の**「4領域同時再設計」**を先行公開する**実験的プレビューモデル**（単なる3.8の小型高速版ではない）。

The Paradox (アーキテクチャの矛盾)

総パラメータ
約180B
（N-gramやMTPを含む巨大サイズ）

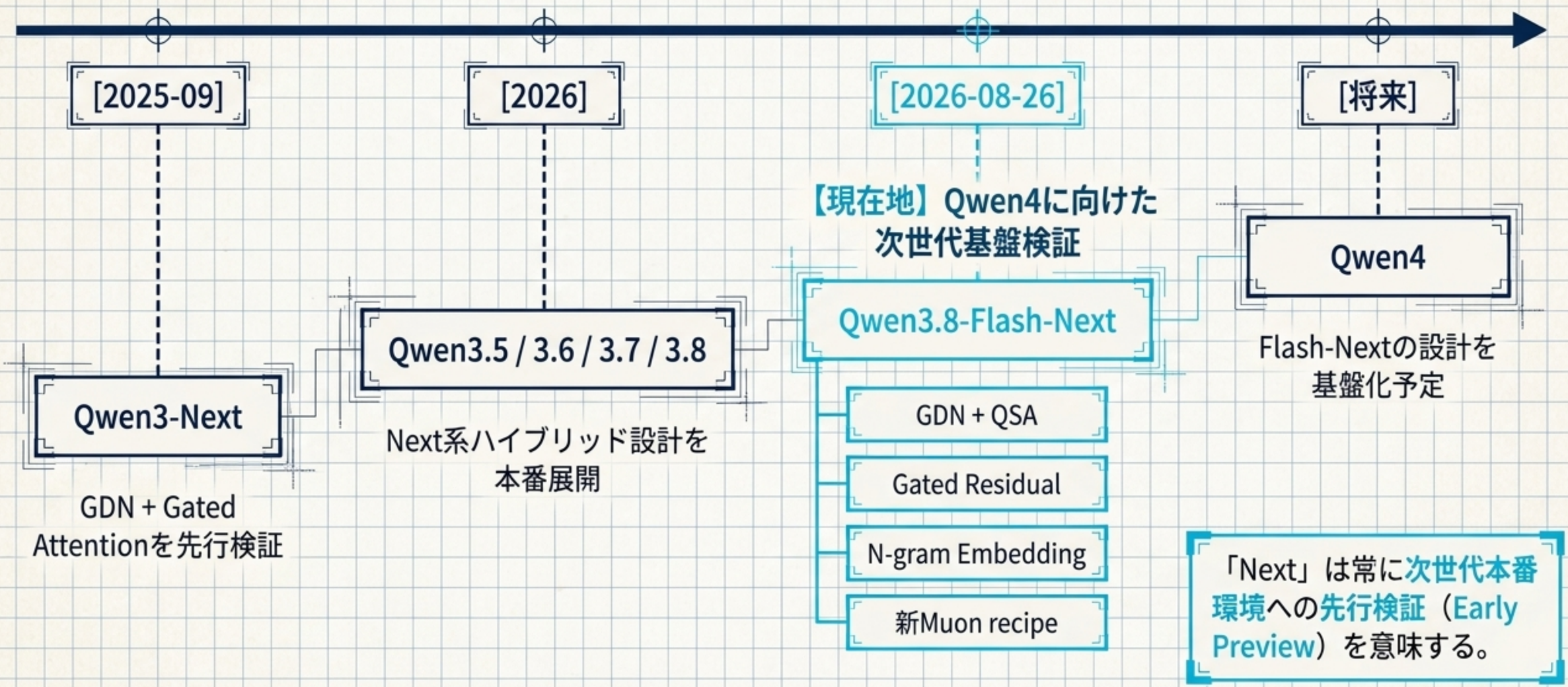
Vs

アクティブ演算
6B
（1トークンあたりの極めて疎なMoE計算）

The Verdict (実務上の評価)

エージェントやコーディング用途には驚異的な性能を発揮するが、「6B並みに軽い」という**誤解は危険**。**巨大なRAMオフロード要件**と、独自の**商用ライセンス制限**に要注意。

Qwen進化の系譜と「Flash-Next」の戦略的位置づけ

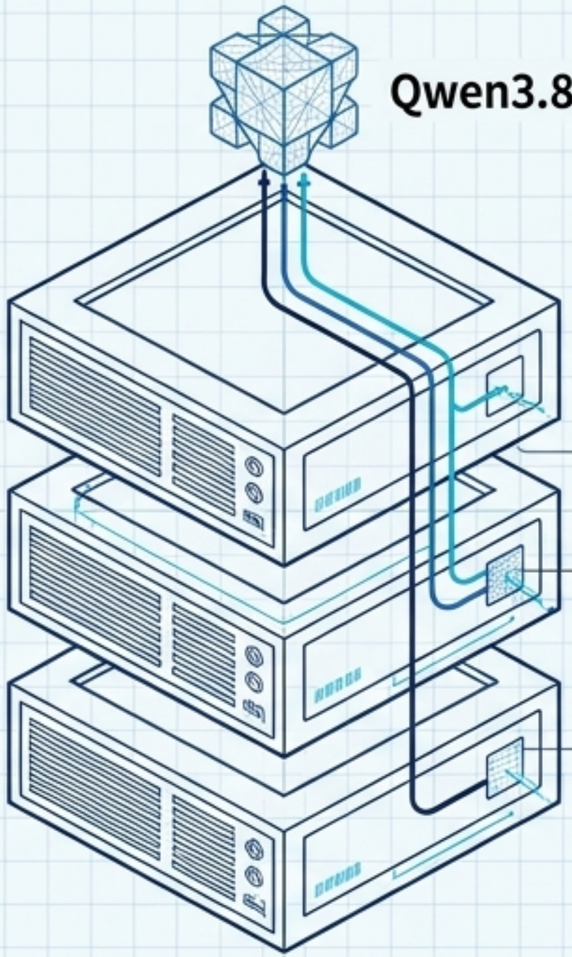


Alibaba AIスタックにおけるQwen3.8-Flash-Nextの提供形態

Layer 1: 研究・自社ホスト層
(Open weights)

Layer 2: 本番API層
(Cloud Production)

Layer 3: アプリケーション層
(End-user/Agent)



Qwen3.8-Flash-Next

- **Hugging Face**
  セルフホスト用
- **ModelScope**
  セルフホスト用
- **QwenCloud production版**
  OpenAI / Anthropic互換API、
  1Mコンテキスト標準、ビルトインツール
- **QwenWork Standard**
  Qwen Code / Agent workloads

⚠ 「Open-weightモデル」と「Production版」は混同注意。
API版には自動セーフティや1M標準対応などの運用仕様が追加されている。

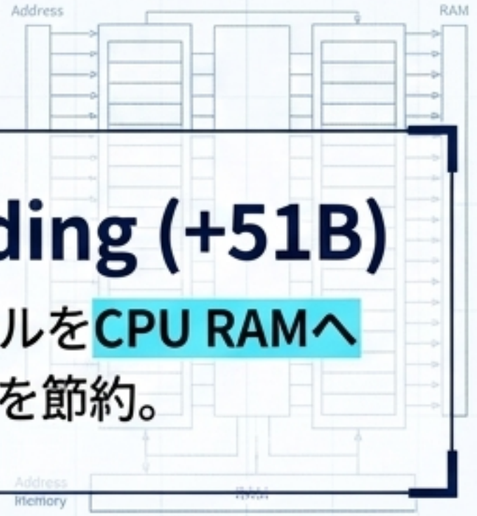
Qwen4をプレビューする「4つの技術的柱」

Attention（注意機構）

GDN + QSA

全トークンを見るのではなく、**マイクロブロック単位**で重要な領域だけを**疎（Sparse）**に選択。

Memory（記憶領域）



N-gram Embedding (+51B)

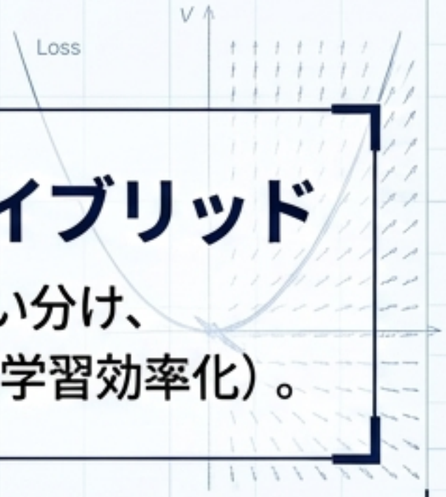
巨大なルックアップテーブルを**CPU RAMへ非同期オフロード**し、GPUを節約。

Architecture（残差接続）

4–branch Gated Residual

従来の単一経路から**4分岐に拡張**し、
層間の情報フローと学習の安定性を最適化。

Optimization（学習手法）



Muon / AdamW ハイブリッド

重みの性質で最適化手法を使い分け、
バッチウォームアップを廃止（学習効率化）。

Compute-Sparse + Memory-Rich：計算を削り、記憶をRAMへ逃がす

[総パラメータ 約180B]

GPU（計算処理）

【Active計算: 6B】

Main LM（125B）

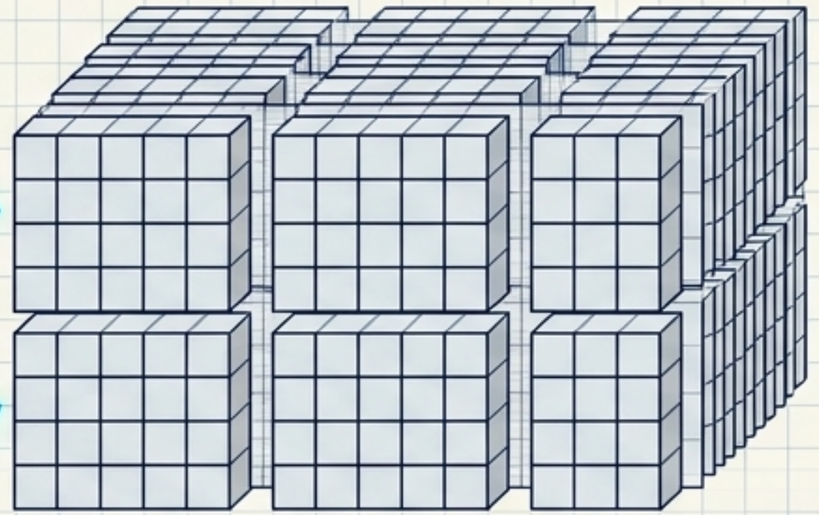
512のExpertsのうち「10 routed + 1 shared」のみが活性化。

MTP (4B)

1層追加のマルチトークン予測。

非同期オフロード

CPU Host RAM（記憶領域）



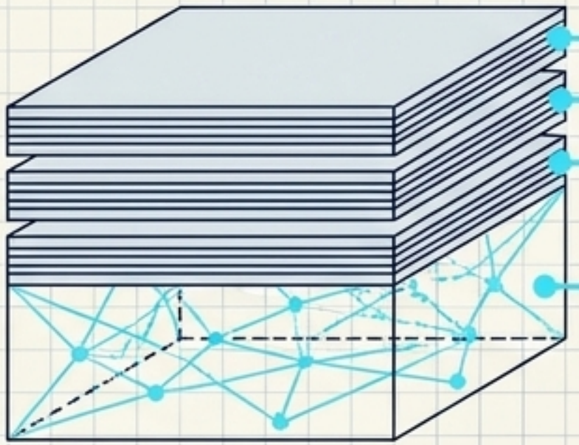
N-gram Memory (51B)

巨大なルックアップ辞書（bi/tri-gram）。
乗算量が小さくGPU計算と並列化可能。

パラダイムシフト：演算（Compute）は極限まで軽くし、モデル容量（Capacity）はメモリ中心に拡張する新設計。

長文コンテキストの制覇：GDNとQSAのハイブリッド構造

全48層は [GDN × 3] + [QSA × 1] の12回リピート構造。



GDN (圧縮状態)：
過去の履歴を固定的・圧縮的な状態 (state) に畳み込み、毎層の重いAttention計算を回避。

QSA (グローバル検索)：
4層に1回、重要なコンテキスト (マイクロブロック単位) だけを軽量インデクサがSparseに選択。

【1Mトークン条件時】

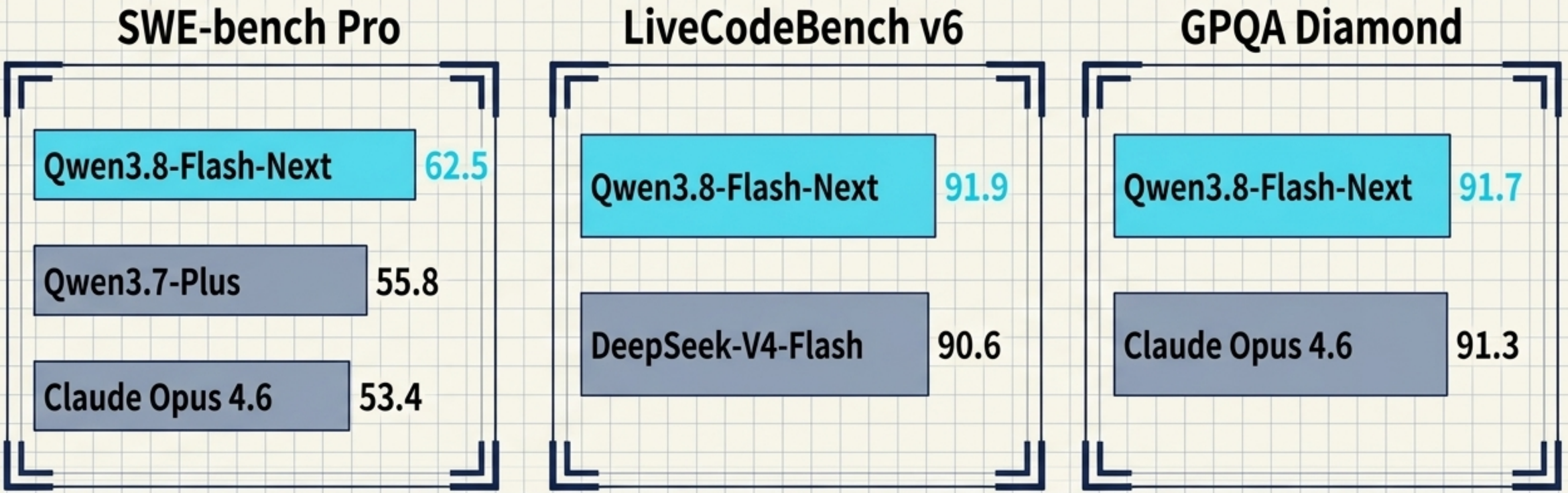
Attention Kernel 高速化：

Prefill 最大 **7.6倍** / Decode 最大 **4.9倍**

プレフィックスキャッシュヒット時 (90%)：

Qwen3.7-Plus比で **スループット 8.6倍**

性能の現実①：エージェントとコーディングの圧倒的エンジン



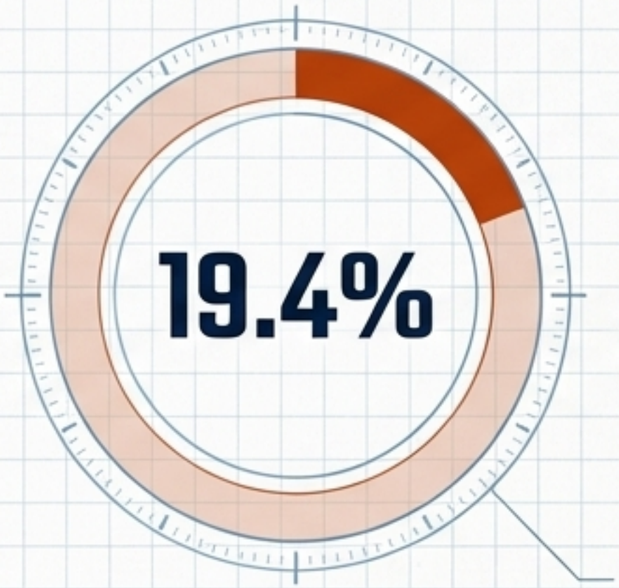
「6B Active」という極小の計算量でありながら、数十万トークンのコードベースやツール履歴（Tool trace）を保持するAgentワークロードで真価を発揮する。

性能の現実②：万能ではない「弱点」の把握

競合への敗北領域

NL2Repo-Bench	Qwen3.8-Flash-Next 48.1	VS	DeepSeek-V4-Flash 54.2
HLE	Qwen3.8-Flash-Next 35.9	VS	Claude Opus 4.6 40.0

実世界GUI操作の壁



OSWorld 2.0 Binary (完全成功率): 19.4%

「全競合を凌駕した」という単純な解釈は危険。ベンチマークの数ポイント差は決定的ではなく、特に自律的なGUI/OS操作には依然として人間の確認（Human-in-the-loop）が必須。

導入インフラの現実：「6Bの軽さ」に騙されてはいけない

演算量 (Compute - Light)

- Active 6B: 演算速度は極めて高速。
- Cloud API: Input $0.16 / Output $0.47 (per 1M tokens) と非常に安価。

メモリ (Memory - Heavy)

- Weight Size: BF16 約360GB / FP8 約185GB / GGUF(量子化) 約72.5〜111.3GB。
- Hardware Needs: 10〜20GB級のコンシューマーGPUには収まらない。
- N-gramオフロードのため高速なHost MemoryとPCIe/NVLink帯域が絶対条件。

Cloud APIの安さと、オープンウェイトを自社ホストする際のTCO（総所有コスト）を混同してはならない。

監査を阻む「ブラックボックス」要件

Unconfirmed Elements（レッドチーム/監査リスト）

UNCONFIRMED 学習コーパス総数・データソース・混合比 [未確認]

UNCONFIRMED Knowledge Cutoff（知識のカットオフ時期） [未確認]

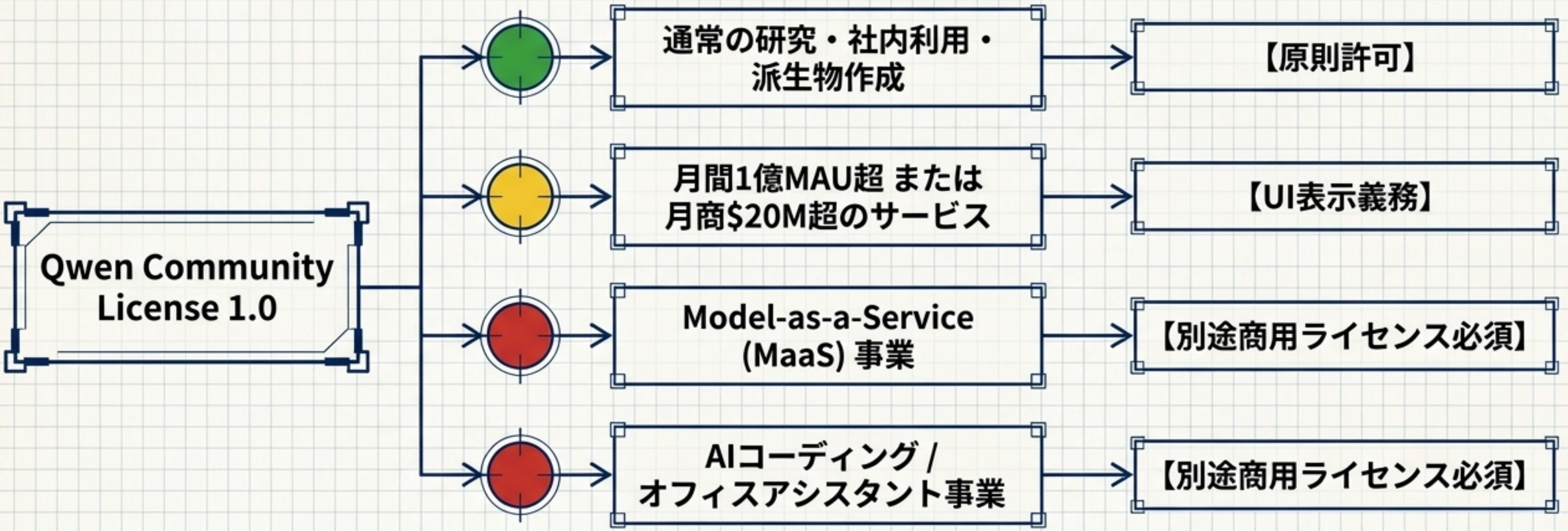
UNCONFIRMED 汚染対策（Contamination deduplication） [未確認]

UNCONFIRMED 安全性評価の詳細（Red-team, Jailbreak検証等） [未確認]

戦略的アドバイス（Strategic Advice）

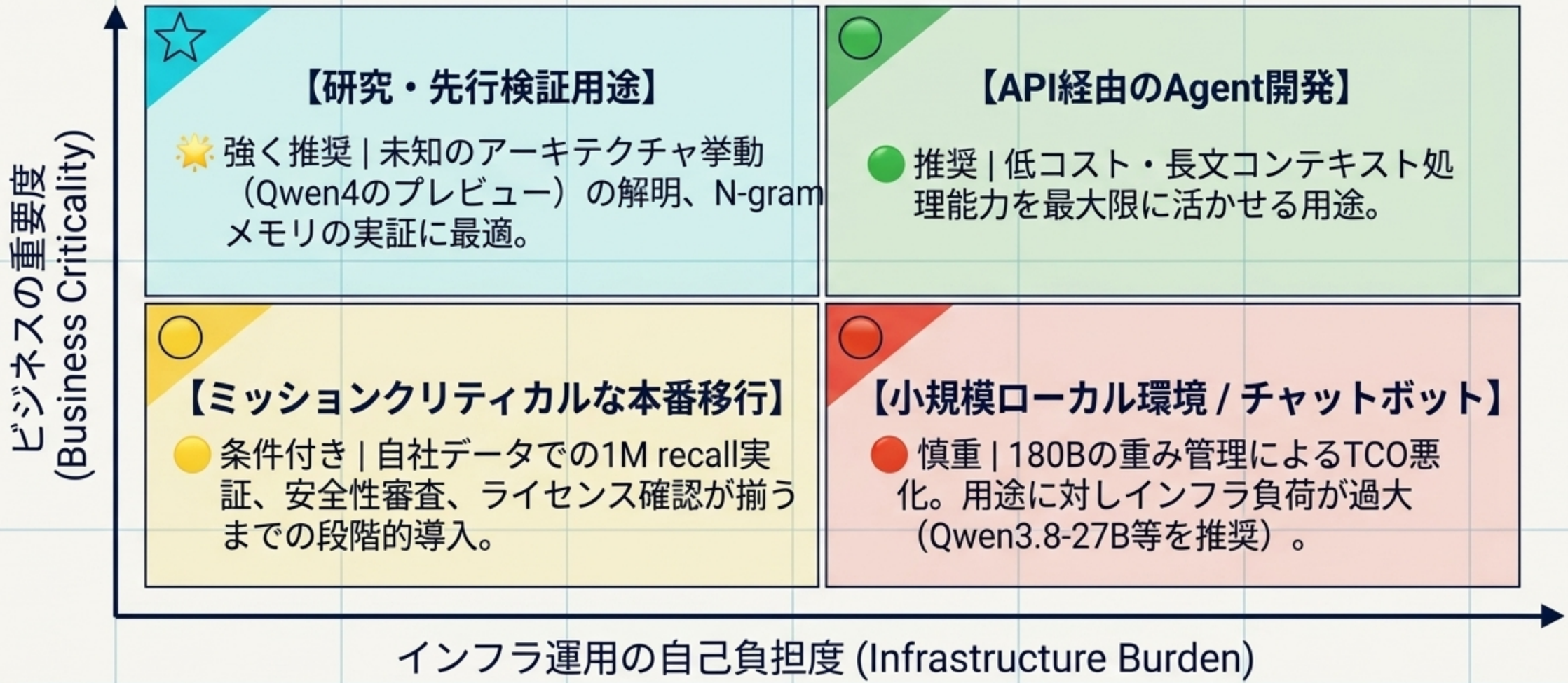
QwenCloud（API）側には自動セーフティガードレールが存在するが、
ダウンロードした重み（Open weights）自体に強制モデレーションはない。
セルフホスト時の安全管理は導入企業の**完全な自己責任**となる。

法務リスクチェック：無条件の「オープンソース」ではない



競合するSaaS（コーディングエージェント等）をこのモデルで構築する場合、
通常のオープンソースモデルとは全く異なる法務審査が初期段階で不可欠。

ユースケース別 導入推奨マトリクス



結論：Qwen3.8-Flash-Nextが示すAIの未来図

Paradigm Shift (設計思想の転換)

パラメータを単純に増やす時代からの脱却。「**Compute-Sparse**（計算の極小化）+ **Memory-Rich**（記憶のRAM退避）」が次世代LLMのトレンドになる強力な実証実験。

Hyper-Specialization (用途の特化)

万能なチャットボットではなく、数十万トークンのツール履歴やコードを処理する「**Agentic / Coding**」用途の特化型エンジン。

The Deployment Rule (導入の鉄則)

カタログスペック（「8.6倍高速」「6Bで軽い」など）を鵜呑みにせず、自社のインフラ帯域（TCO）と法務要件と照らし合わせた厳密な**PoC**が不可欠。

Appendix & Source Materials

	QwenLM/Qwen3.8-Flash-Next GitHub Repository (Technical Report, Code)
	Qwen3.8-Flash-Next Hugging Face Model Card (Weights, Metrics)
	QwenCloud Documentation (API Specs, Safety Guidelines, Pricing)
	Qwen Community License 1.0 (Hugging Face / GitHub)
	PC Watch Article: 「Qwen3.8-Flash-Next」無償公開、Opus 4.6匹敵… (Local VRAM sizing context)

All data extracted as of report compilation date. Third-party independent verification is recommended for critical production deployment.