

[DATE] 2026.08.27 ●

[CLASSIFICATION] Strategic Intelligence Briefing

[TARGET] 経営層・R&D・知財部門

Qwen3.8-Flash-Next 調査レポート

2026年8月公開：事実確認、技術的特徴、そして日本企業の実務への示唆

経営層が把握すべき3つの重要ファクト



1. Qwen4アーキテクチャの早期プレビュー

- 総パラメータ180B（主モデル125B）
- アクティブ6BのマルチモーダルMoE
- 次世代「Qwen4」の技術仕様を先行実装



2. 劇的なコスト削減とOpus 4.6超えの主張

- 前世代比で学習コストを約9分の1に削減
- コーディングや長文脈タスクでの圧倒的性能
- 自社ベンチでClaude Opus 4.6を凌駕と主張



3. 性能の未検証性と独自の商用ライセンス

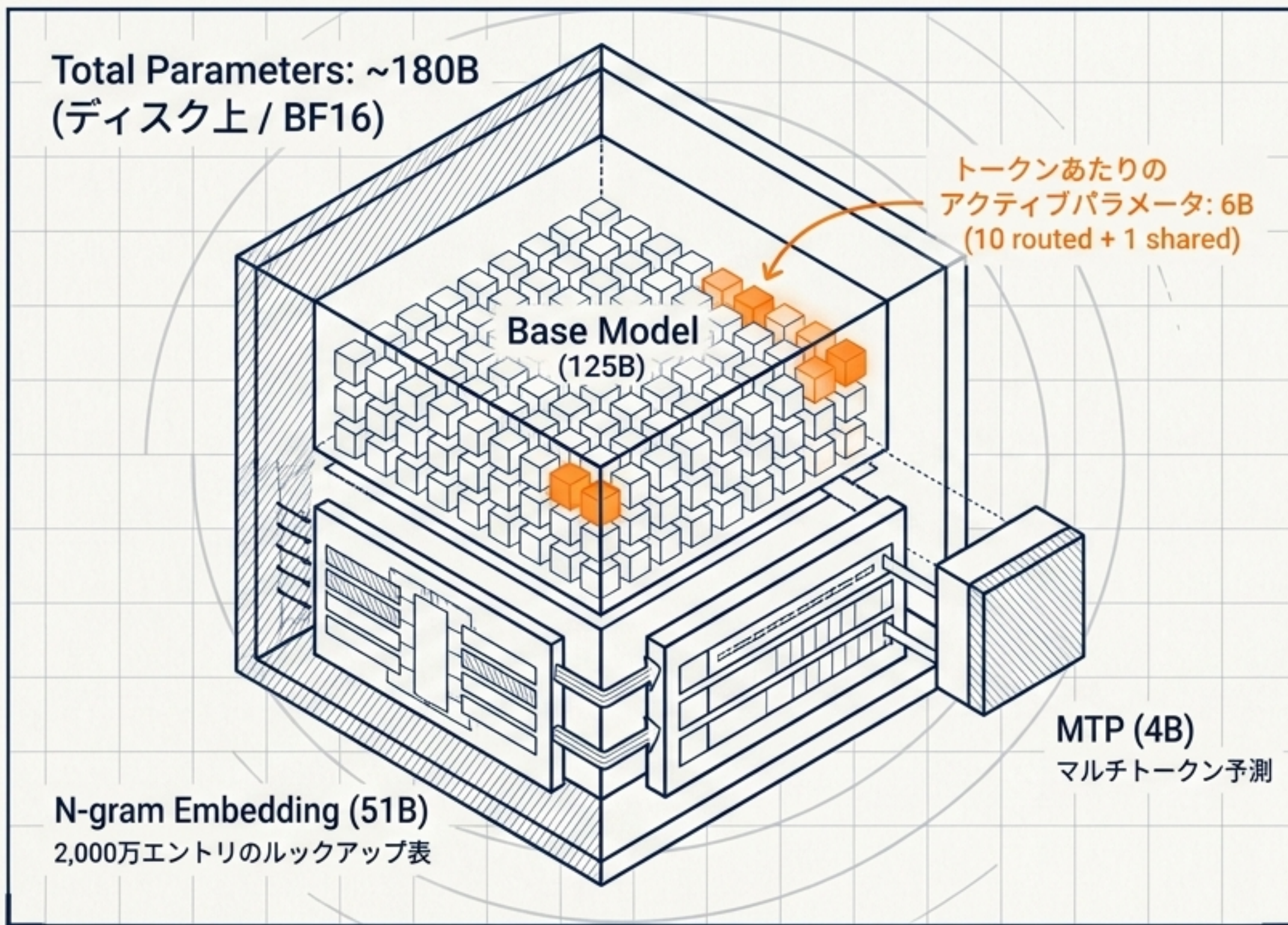
- 独立した第三者機関による評価は未了
- Apache 2.0ではない独自ライセンス
- 「Qwen Community License 1.0」を採用

近縁モデルの整理と曖昧さ回避

モデル名称	公開形態	規模/アーキテクチャ	役割と位置づけ
Qwen3.8-Flash-Next 対象	オープンウェイト	180B / MoE	本番前のアーキテクチャ実験・コミュニティ検証用
Qwen3.8-Flash	API提供のみ	非公開	同日（8/26）ローンチされた量産版。入力\$0.16/1M Tokenの低価格。
Qwen3.8-Max	API / オープン	2.4兆 (Active 95B)	8/3公開の最上位フラグシップモデル。
Qwen3-Next	オープンウェイト	80B / MoE	2025年9月公開の前世代レビュー（旧モデル）。

⚠ 誤称注意: 一部で流布している「~~Qwen3.8 Next-Flash~~」という名称は誤り。

モデル規模と基本アーキテクチャの解剖



基本スペック (Specs Sidebar)

- Context Length:
ネイティブ 262,144 トークン
(YaRN拡張で最大 1,000,000 トークン)
- Layers:
48層
- Layout:
3x(Gated DeltaNet) に対し
1x(QSA) の4対1レイアウト
- Checkpoint:
単一モデルで機能。Instruct /
Thinking の切り替え可能

Qwen4を支える4つのアーキテクチャ刷新

※Qwen3.7-Plus比で「学習コスト1/9」を実現

1. Hybrid Attention (QSA)

Mechanism: 軽量Indexerによる重要コンテキストの選択。

Impact: 1Mトークン時のprefill最大7.6倍高速化、スループット8.6倍。

2. Gated Residual

Mechanism: 4分岐の残差ストリームをデータ依存ゲートで制御。

Impact: 学習の安定性と表現力の両立。

3. N-gram Embedding

Mechanism: 第2層に配置された51Bの巨大なルックアップ表。

Impact: 演算量を増やさず容量拡張。ホストシステムRAMへの非同期オフロードが可能。

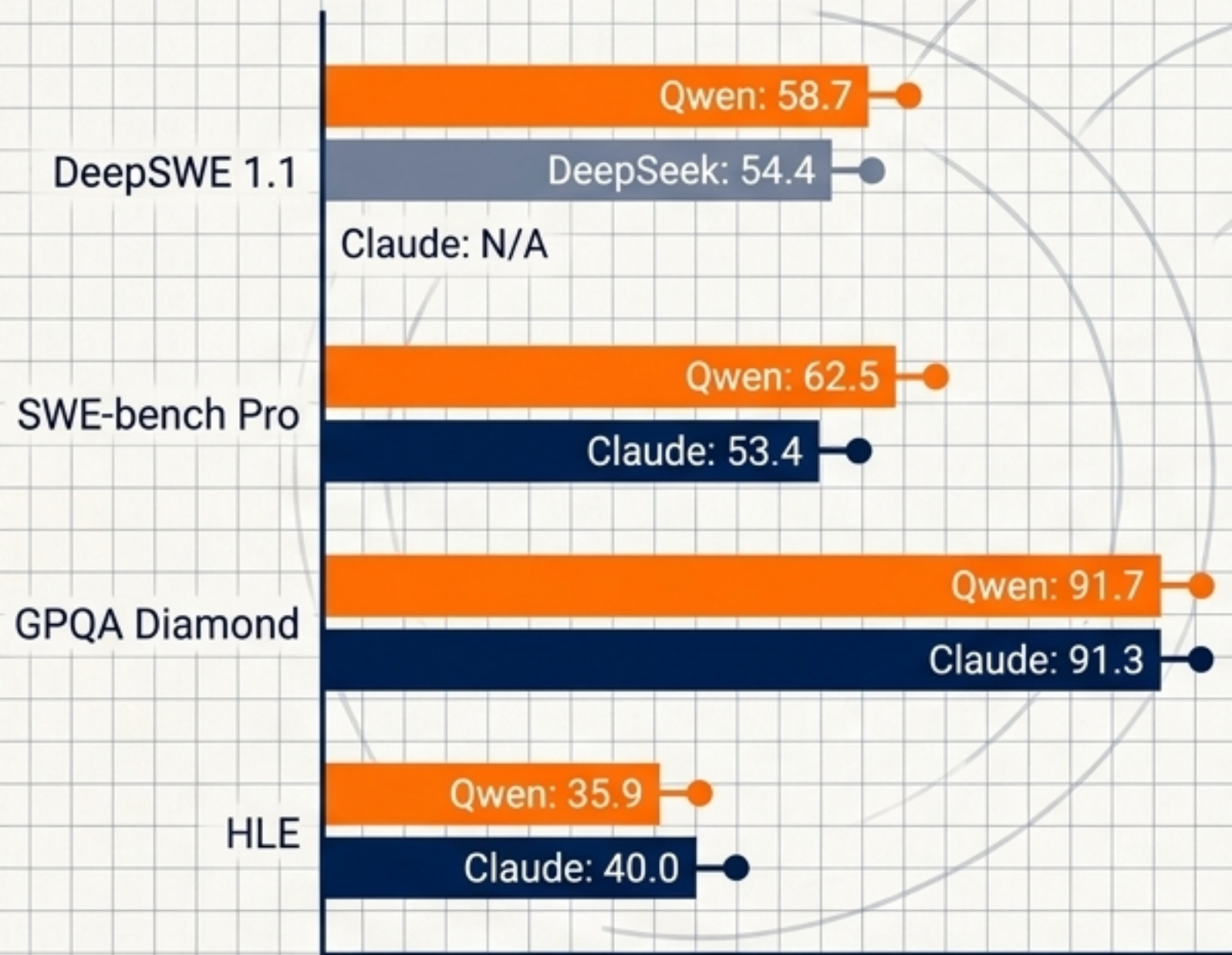
4. Tailored Training

Mechanism: MuonとAdamWのハイブリッド適用。Batch-size warmupの廃止。

Impact: 学習効率の極大化。学習コスト大幅減の主因。

性能評価ダッシュボード：主張 vs. 現実

自社発表スコア (Claims)



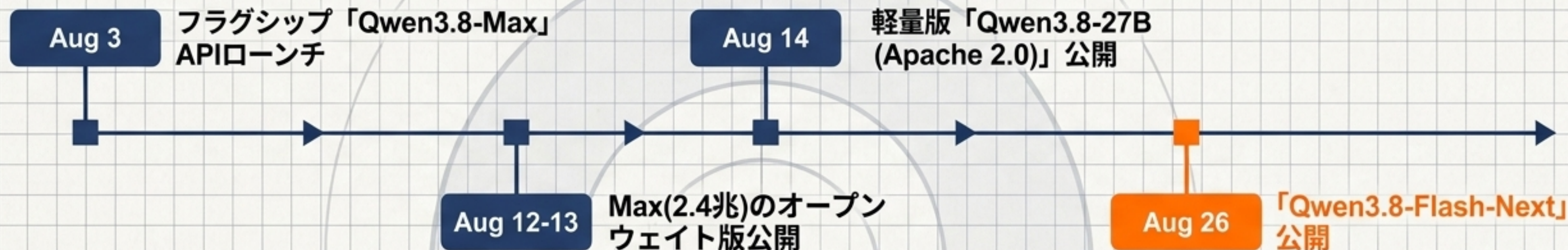
第三者検証ステータス (Reality)

Leaderboard	Qwen Rank	Claude Rank
Artificial Analysis		
LMarena		
SWE-bench		

⚠ Caveats (留保事項)

- 比較対象が最新のOpus 4.8ではなく、Opus 4.6 (26年2月) である。
- 検証スコアの一部に自社製ベンチマーク (CoWorkBench等) が含まれる。

Strategic Blitz: Alibabaの市場覇権シナリオ



The DeepSeek Playbook

クローズドAPIへの依存から、「オープンウェイト+超低価格」による市場のコモディティ化へ完全な戦略転換。

(API量産版はMaxの1/12の価格水準)

Why release 'Next'?

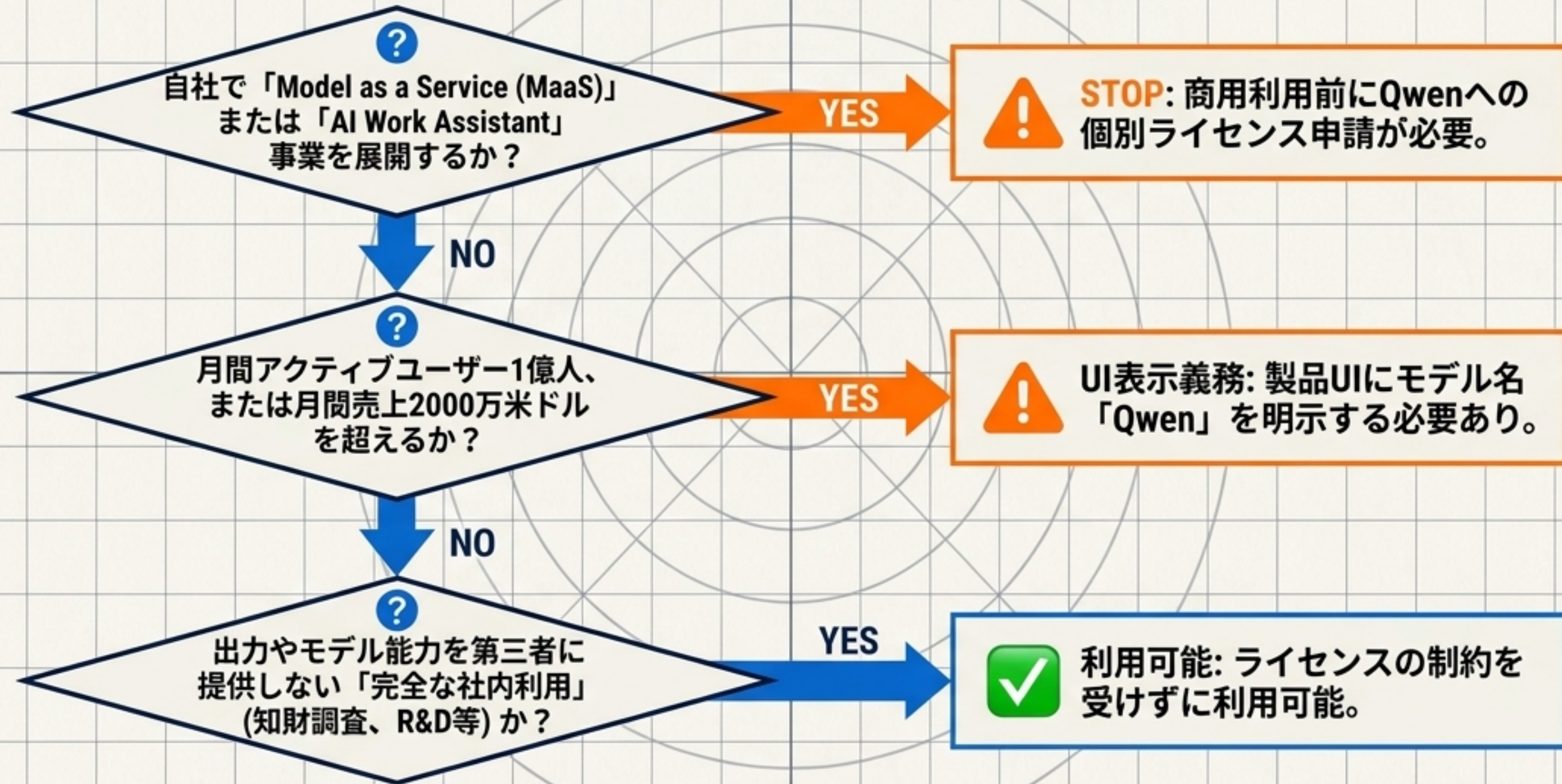
Qwen4の完成を待たずアーキテクチャを先行公開する複合的意図：

1. ランタイム・量子化ツール等のエコシステム早期構築
2. コミュニティ環境での新アーキテクチャの大規模グレーテスト

ライセンス判定ツリー：qwen-community-1.0の法務審査

NOT Apache 2.0

NOT OSI Open Source



※噂された「特定地域(米国・EU等)での禁止」条項はライセンス本文に存在しない誤情報である。

ローカルインフラ要件とアーキテクチャの優位性

メモリ要件の課題 (Memory Requirements Challenge)

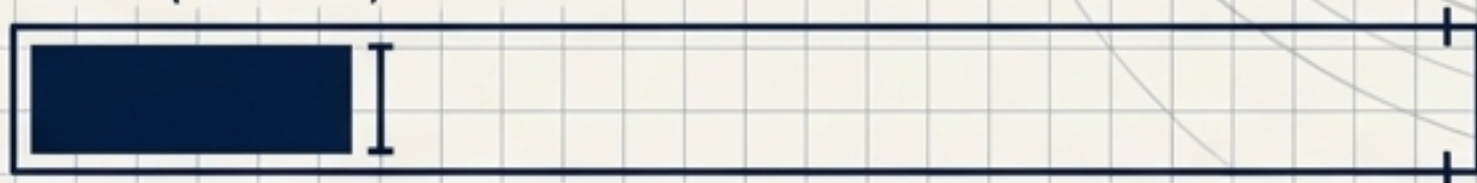
BF16 (Full): ~355GB (非現実的)



4-bit 量子化: 80~90GB (複数GPU要)



1-bit (Unsloth): ~75GB



N-gramソリューション (The N-gram Solution)



The N-gram Advantage

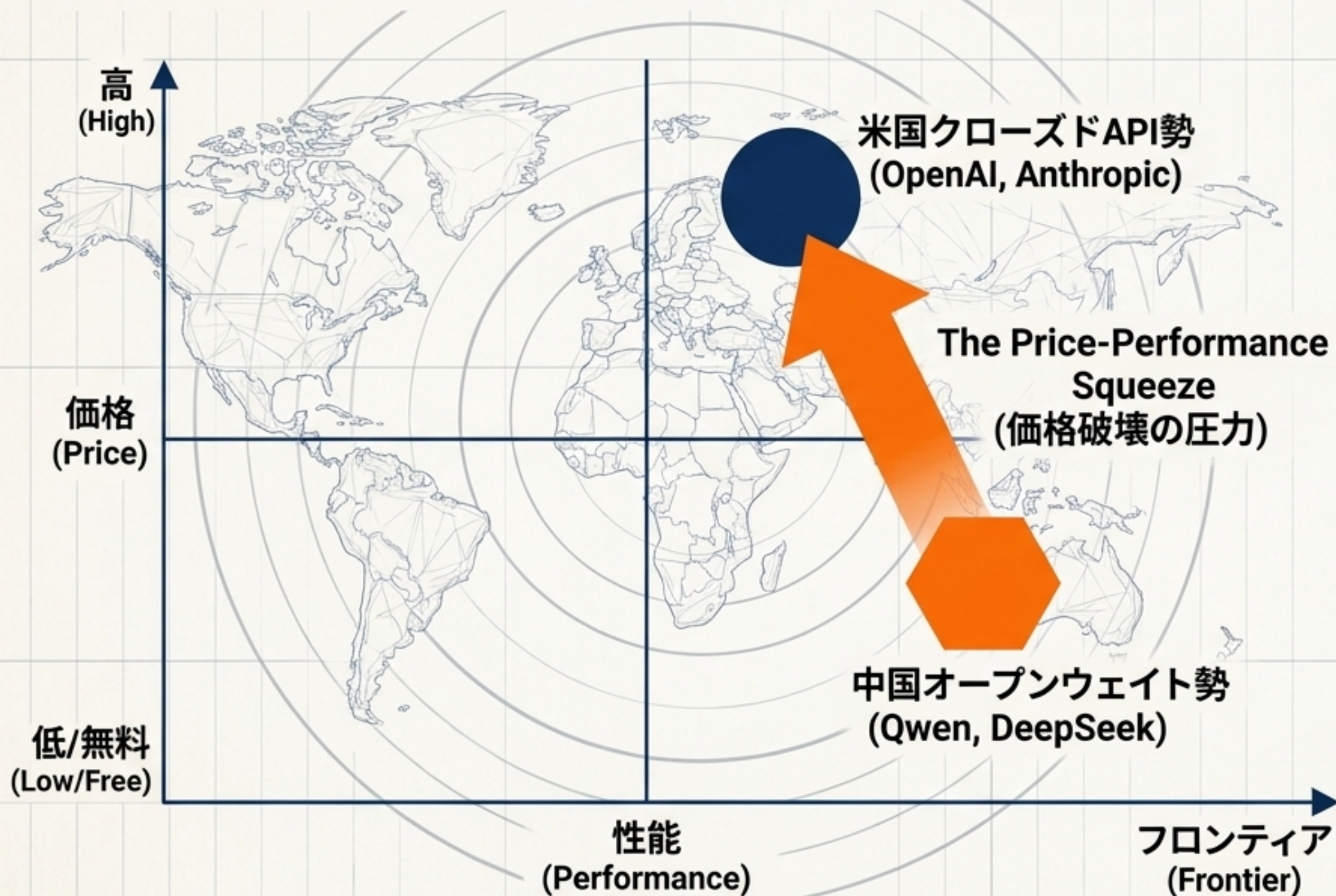
- 巨大なN-gram表をホストシステムRAMに非同期オフロード可能な設計。
- GPU VRAMが潤沢でなくとも、高RAM構成の自社サーバーで動作させやすい。

実務への示唆: 機密データ（特許明細書・未公開R&D）を外部に出せない知財・R&D部門のVPC/ローカル隔離運用において、極めて強力な選択肢となる。

日本企業向け導入・検証ロードマップ



The Global Implication: 米中フロンティア競争の構造変化



マクロ分析

1. **価格圧力: フロンティア級一性能に肉薄しつつ価格を1桁下げる戦略により、米国勢の高価格維持を持続困難にしている。**
2. **規制回避 (Regulatory Bypass):** 米国の政策枠組みにおいて「**オープンウェイト**」が連邦レビュー対象外とされる隙を突き、グローバル覇権を握る構造。

評価の留保事項と情報の限界 (Caveats)

□	自社公表値への依存: 性能数値や普及規模(30億DL 等)はAlibabaの自己申告であり、第三者実測値との乖離が存在する。
□	比較対象の非対称性: Claudeとの比較は「最新の Opus 4.8 」ではなく、2026年2月時点の「 Opus 4.6 」に対するものである。
□	パラメーターの表記揺れ: 「 176B 」との報道もあるが、本資料では公式内訳に基づく「 180B 」を採用している。
□	法的助言の非該当: 本資料のライセンス解釈は法的助言ではない。 MaaS 該当性の判断等には自社法務の確認を要する。

➤➤➤ 結論：アーキテクチャのブレークスルーが独立検証されるまで、
本格的な実働環境への全面移行は保留すべきである。 <<<