[DOCUMENT_CLASSIFICATION: STRATEGIC EVALUATION]

[TARGET: ENTERPRISE AI ARCHITECTURE]

[UPDATE: 2026-04-26 JST]

中国LLM最前線：エンタープライズ導入・評価ドシエ

4大最新モデル（DeepSeek, Kimi, Qwen, MiMo）の絶対性能、導入コスト、透明性の完全解剖

[DeepSeek-V4-Pro] | [Kimi K2.6] | [Qwen3.6-27B] | [MiMo-V2.5-Pro]

結論：4大モデルの実力と「隠れたコスト」

DeepSeek-V4-Pro

［至高の長文処理］

1Mコンテキストと推論効率で他を圧倒するが、運用には巨大な計算資源が必要。

Kimi K2.6

［最強のエージェント］

コーディングと長時間自律実行でトップ。ただしモデルが巨大で情報開示が薄い。

Qwen3.6-27B

［最高の実務性］

サイズ対性能比が極めて高く、自前運用・日本語チューニングにおける本命。

MiMo-V2.5-Pro

［未知のAPI特化］

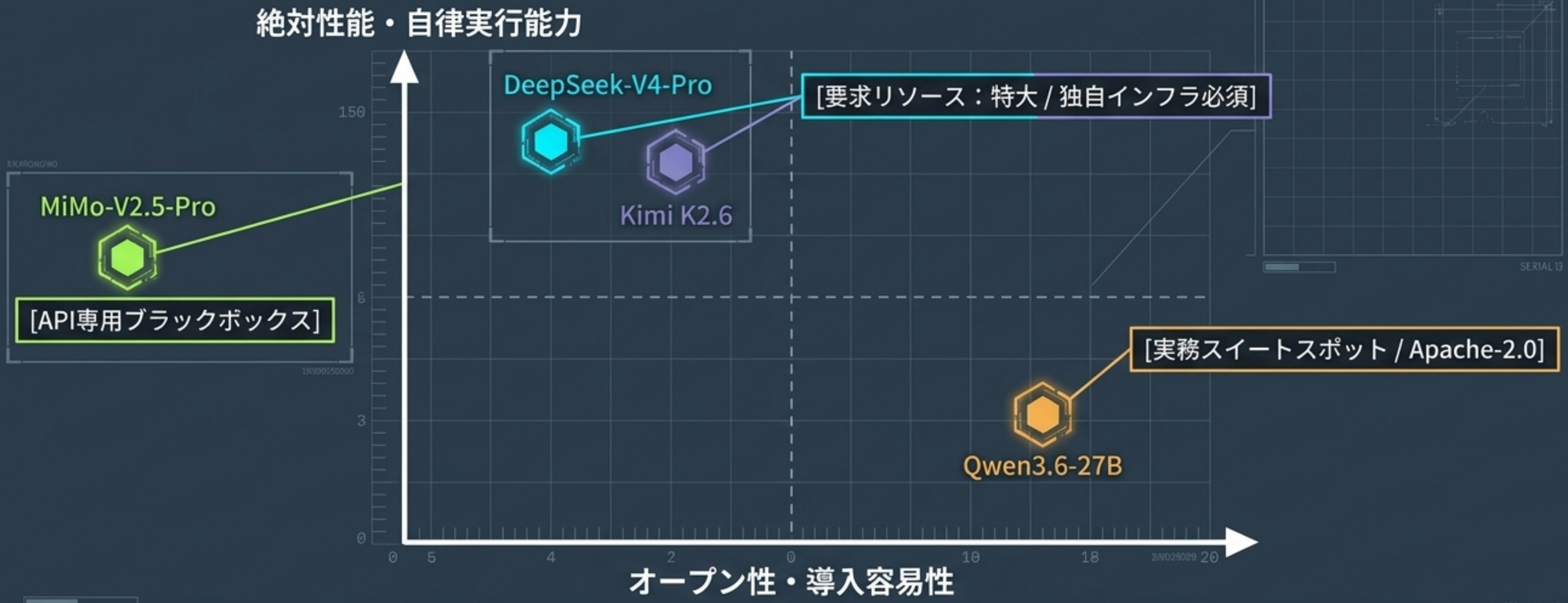
1000回超のツール呼び出しに耐えるが、透明性が最も低くエンタープライズ監査には不向き。

コア・ダイアグノスティック・マトリクス

● 完全装備 / ◑ 部分的 / ○ 不足・未公開

	DeepSeek	Kimi	Qwen	MiMo
基本スペック	1.6T MoE 1M Context	1T MoE 256K Context	27B Dense 262K Context	未公開 (V2比 1T超)
導入容易性	◑ (巨大クラスタ必須)	◑ (高級GPU群必須)	● (TP=8 GPU構成で公式対応)	○ (API専用・セルフホスト不可)
ライセンス	MIT (商用可)	Modified MIT	Apache-2.0 (商用最適)	未公開 (規約依存)
自律エージェント能力	●	● (HLE 54.0)	◑	● (APIベースで最高峰)
透明性・安全性	◑ (データ詳細未公開)	○ (System card未確認)	◑ (Post-training公開済)	○ (完全ブラックボックス)

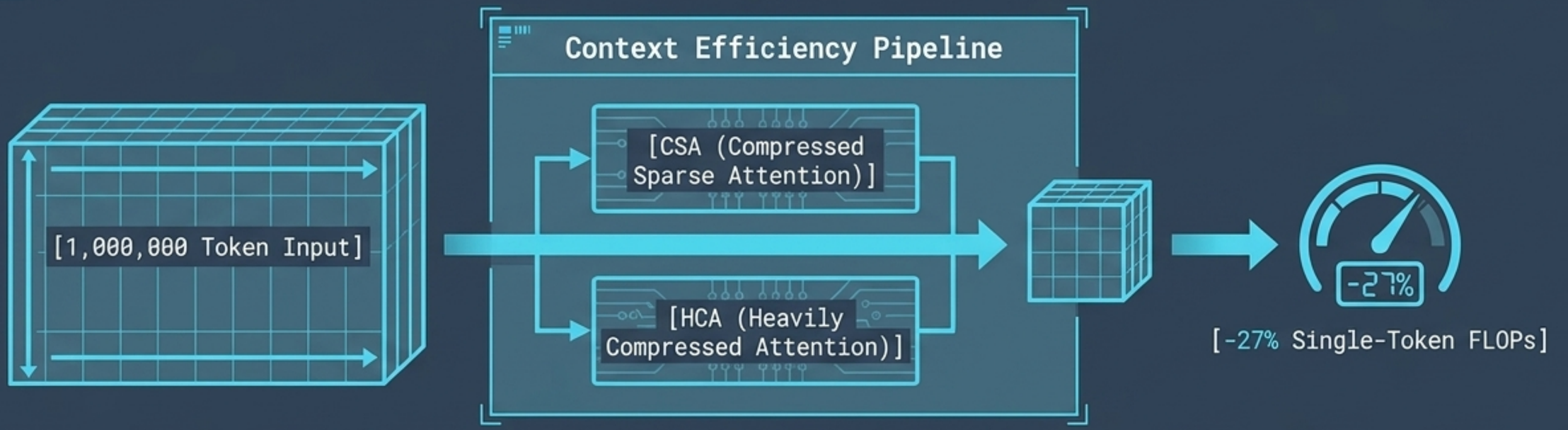
デプロイメントの現実：性能 vs. 運用摩擦



「性能の絶対値」を追うと運用ハードルと監査リスクが跳ね上がる。自社インフラでの運用が必須な場合、Qwen3.6-27Bが唯一の現実的選択肢となる。

[1.6T MoE] [1M Context] [MIT License]

DeepSeek-V4-Pro：長文処理の限界突破



Key Specs

Total/Active Params: 1.6T / 49B
Training Data: 33T Tokens (Pro版)
Benchmarks: MMLU-Pro 87.5 / LiveCodeBench 93.5

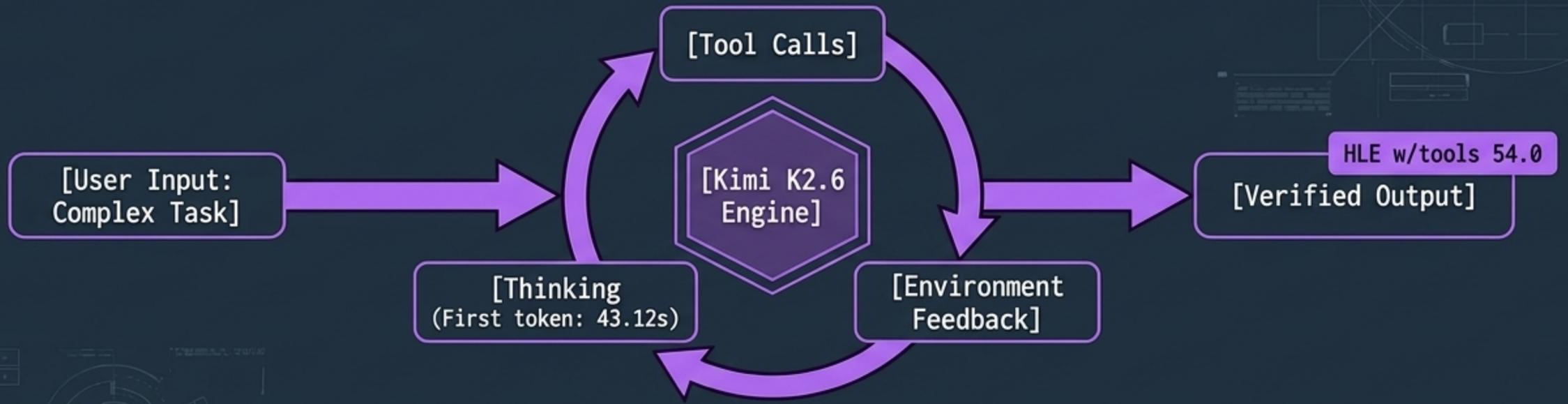
Deployment Reality

API: 36.9 t/s (公式API)
Local: 最低限のGPU構成未公開。実質的にはエンタープライズ級のマルチGPUクラスタが必須。独自のencodingフォルダ方式によるエコシステム統合の手間あり。

[1T MoE] [Modified MIT] [Native Multimodal]

Kimi K2.6：最強の自律型コーディング・エージェント

Agentic Workflow Loop



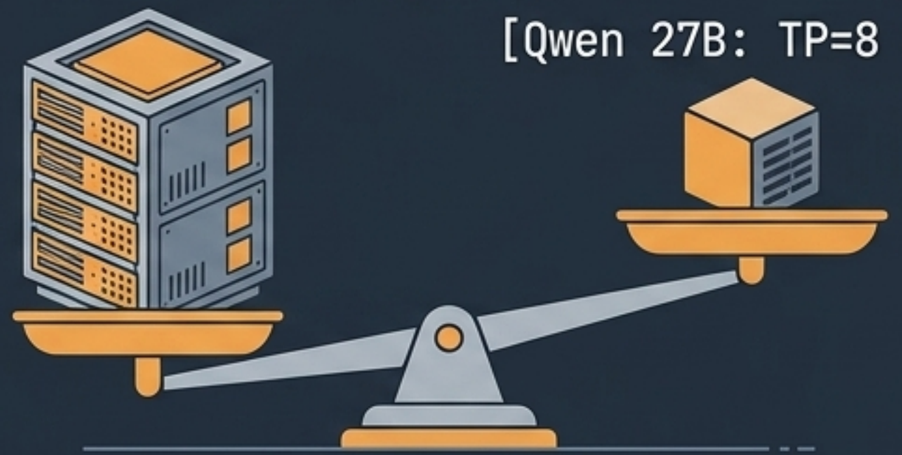
Key Specs	Deployment Reality
Total/Active Params: 1T / 32B Architecture: MLA, MoonViT 400M Benchmarks: BrowseComp 83.2 / SWE Verified 80.2	API: 106.3 t/s（高速）だが推論開始まで数十秒のラグ。 Risk: K2.6固有の学習データ・安全対策の開示が皆無。韓国語で「Anthropic」と名乗る幻覚のコミュニティ報告あり。

[27B Dense] [Apache-2.0] [8 GPU Setup]

Qwen3.6-27B：自前運用・実務導入の最適解

Deployment Weight Scale

[1T+ MoE Server Rack]



[Qwen 27B: TP=8 GPU]

- ✓ [SGLang/vLLM]
- ✓ [llama.cpp]
- ✓ [SFT/DPO Fine-tuning]

Key Specs

Params: 27B Dense（Vision Encoder付き）

Context: 262K Native（最大1.01M拡張）

Benchmarks: SWE Verified 77.2 / GPQA 87.8

Deployment Reality

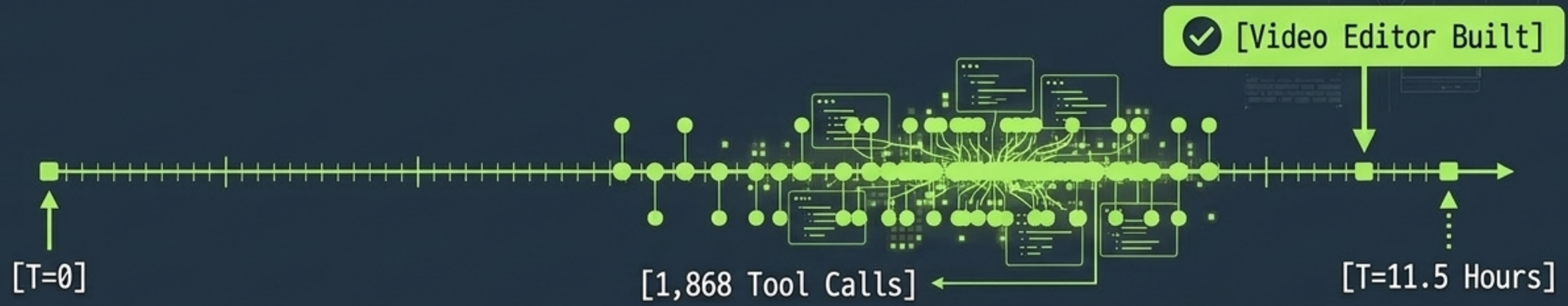
Speed: 62.9 t/s（TTFT 3.88s）

Enterprise Value: 今回唯一のApache-2.0。
自社データでの追加学習（Swift/Llama-Factory）
エコシステムが完全に整備されている。

[Proprietary API] [Agent Scaffold] [Beta]

MiMo-V2.5-Pro：長距離タスク特化のAPIブラックボックス

Endurance Timeline



Token Efficiency: 40–60% fewer tokens than Claude/GPT

Key Specs	Deployment Reality
Params/Data: 完全未公開（V2-Proは1T超） Context: 未公開（長時間維持に特化） Benchmarks: ClawEval 64% Pass^3 / SysY Compiler 233/233	API: 62.0 t/s. AI Studioのみで提供。 Risk: 監査・再現性・法務確認に必要な情報が最も少ない（Safety/System card未確認）。大企業の本番導入には時期尚早。

[SECURITY AUDIT PANEL]

情報的負債（Information Debt）と監査リスク

	DeepSeek	Kimi	Qwen	MiMo
学習データ開示率 (Training Data Disclosure)	⚠️ 量(33T)は開示も、言語分布・独自比率不明。	🚨 K2.6固有の新規データ未公開。	⚠️ 系列(36T/119言語)は開示も、3.6固有は薄い。	🚨 完全非公開。
既知のバグ・安全性 (Known Safety Issues)	⚠️ tool_callsが平文JSONになるAPIバグ。	⚠️ 過剰思考による創造性低下、Anthropic幻覚報告。	⚠️ Safety disclosureが性能情報に比べ薄い。	🚨 Dedicated safety card未確認。

[SYNTHESIS TAKEAWAY]

性能競争は進んでいるが、エンタープライズ水準の「透明性の説明責任」はどのモデルも欠如している。API利用時のデータガバナンス設計が必須。

日本市場戦略：ベンチマークの幻影と現実的アプローチ

⚠「日本語専用ベンチマークの不在」

4モデルすべてが、MMLU-jaやRakuda系の日本専用スコアを公式未公開。

多言語proxyスコア（MMMLU等）は一部あるが、日本市場特有のコンプライアンスや表現のニュアンスは担保されていない。

STRATEGY SHIFT:
From Raw Benchmarks to Practical Architecture

✓「なぜQwenが日本市場で勝つのか」

Qwen

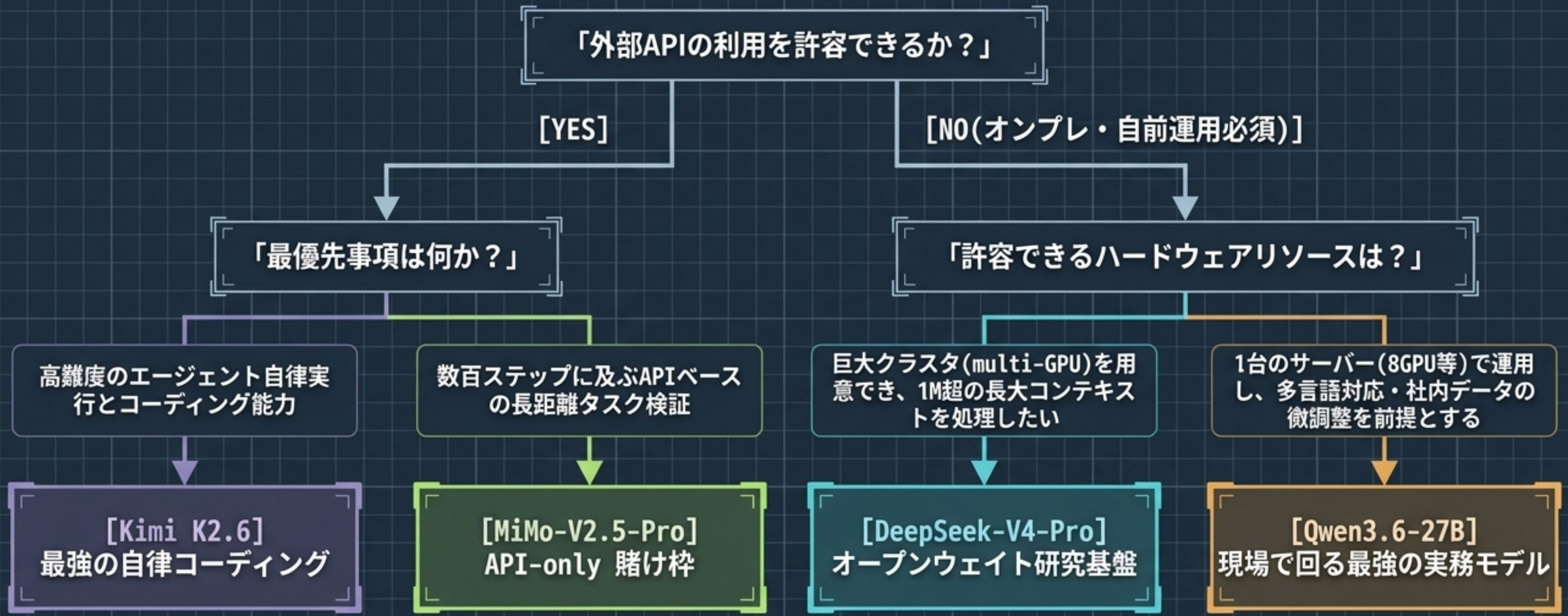
理由1: Qwen系列の201言語・方言対応という強固な多言語事前学習の系譜。

理由2: 27B Denseという「追加学習（SFT/DPO）のしやすさ」。日本語特化の微調整がオンプレミスで容易に完結する。

Qwen

ACTION REQUIRED: 本番採用前に、自社専用の「Japanese Eval Suite（評価基盤）」の構築が絶対に不可欠。

CTO向け：アーキテクチャ選定ディシジョンツリー



Recommendation engine based on capability metrics vs. operational constraints mapped on 2026-04-26. End of Briefing.