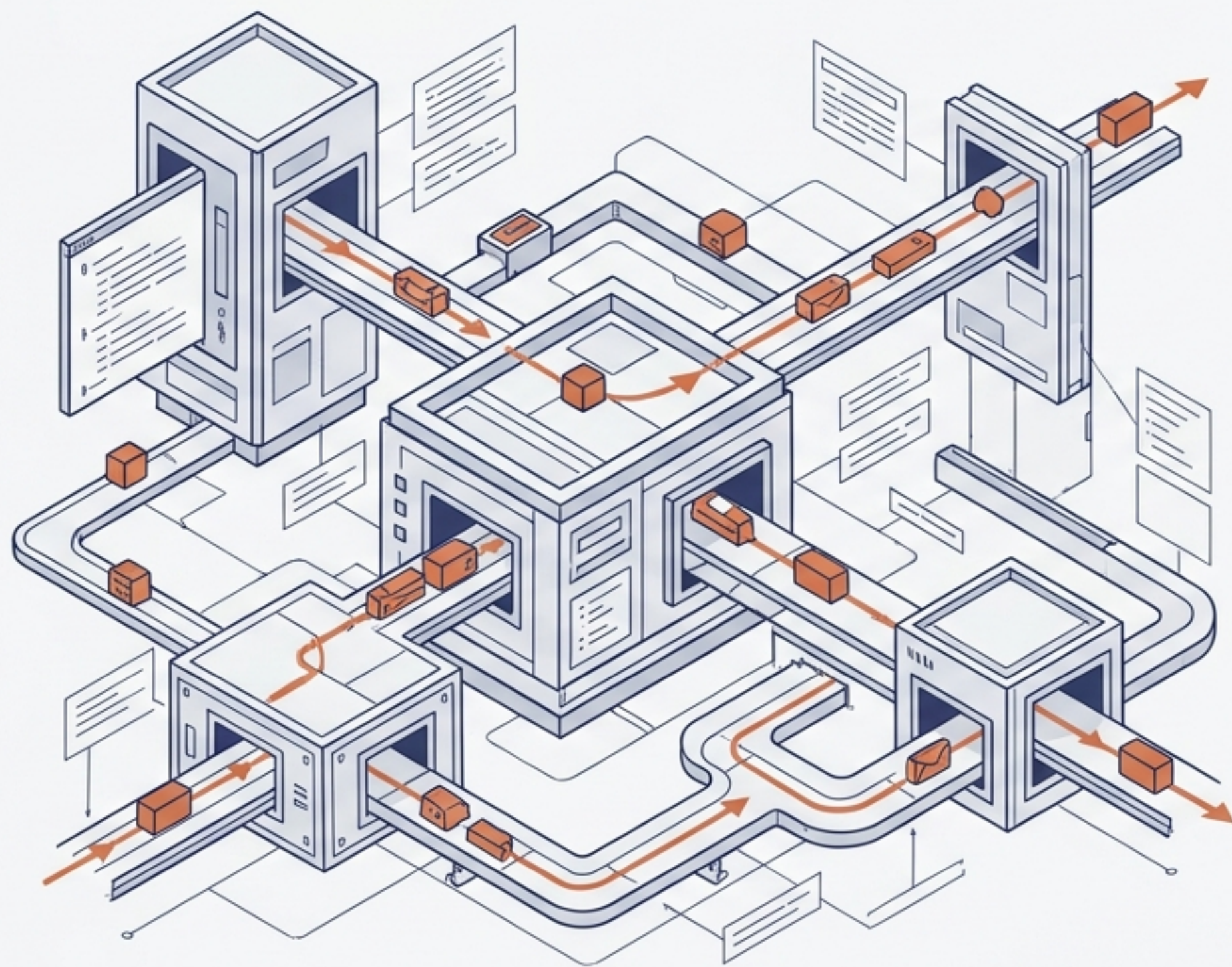


中国AI急進と日本の戦略的転換

コスト破壊、可用性リスク、そして「バーティカルAI」への針路



エグゼクティブ・サマリー

1. 可用性の覚醒

2026年6月の米国製最先端モデル停止が、企業にとっての「能力至上主義」から「可用性・主権重視」への転換点を生んだ。

2. 推論経済性の脅威

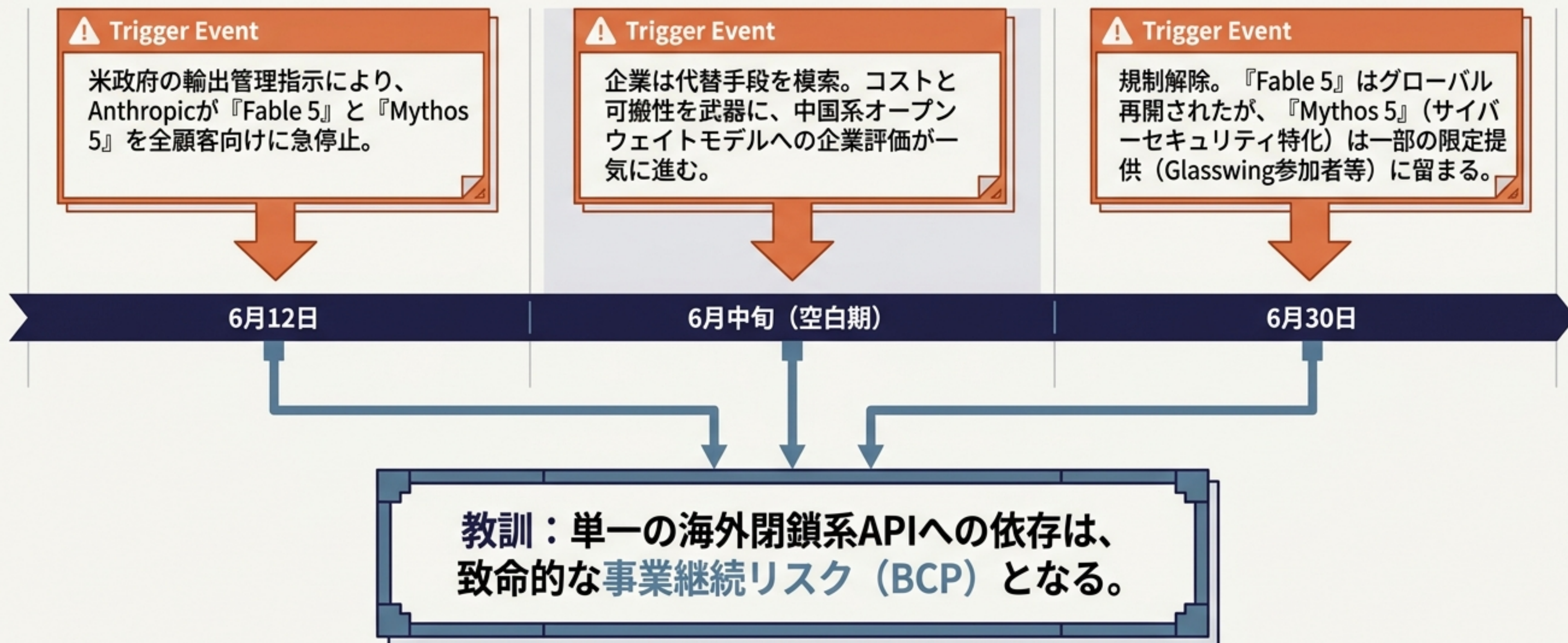
中国モデルの圧倒的低価格単なるダンピングではなく、MoEやMLAなどの技術アーキテクチャに裏打ちされた「設計上の必然」である。

3. 日本の勝ち筋

汎用LLM競争から脱却し、製造・ロボティクス等の現場データ (AI-Ready化) とハイブリッド運用による「バーティカルAI」の覇権を握る。

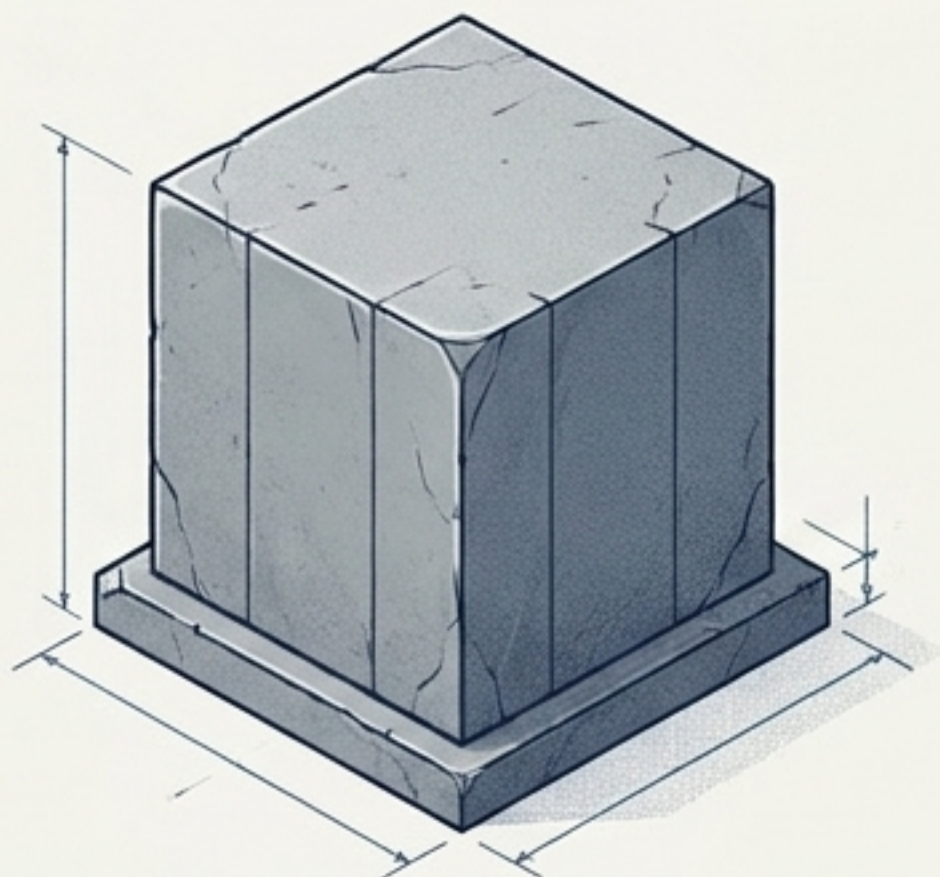
転換点：2026年6月「Anthropicショック」

最先端モデルの「能力」よりも、「政治判断で止まりうるリスク（可用性）」が突如として顕在化した。



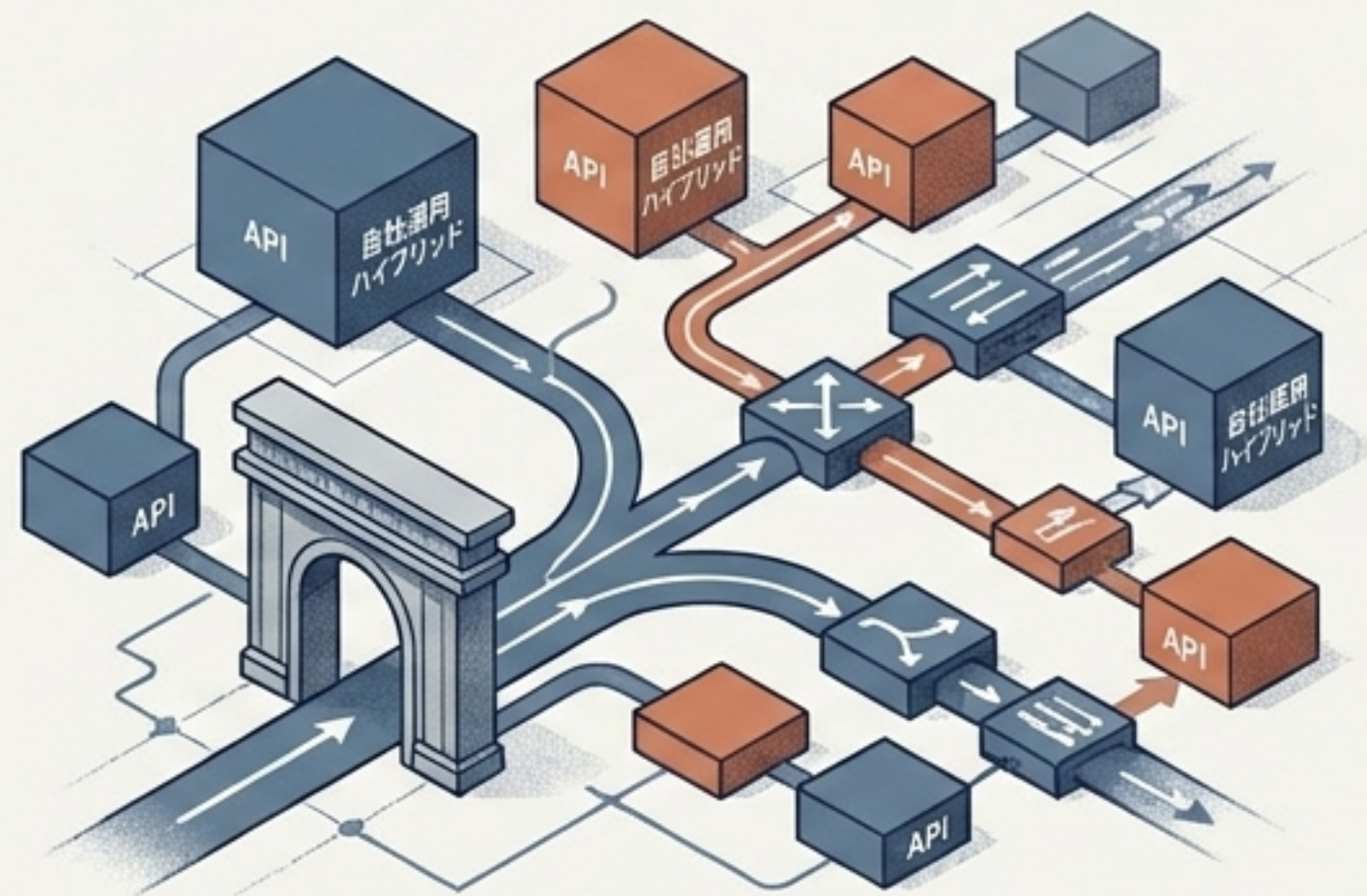
エンタープライズAIの新常識：FinOpsとマルチモデル化

従来のアプローチ



- 評価軸：ベンチマークスコア・絶対的な能力
- 運用形態：単一の最先端フロンティアAPI（海外依存）
- コスト管理：能力至上主義（コスト度外視）

新たなパラダイム



- 評価軸：タスク完了あたりのコスト（Cost-per-task）と移行容易性
- 運用形態：用途別のマルチモデル・ルーティング（API+自社運用ハイブリッド）
- コスト管理：企業FinOpsの論理による厳格なROI管理

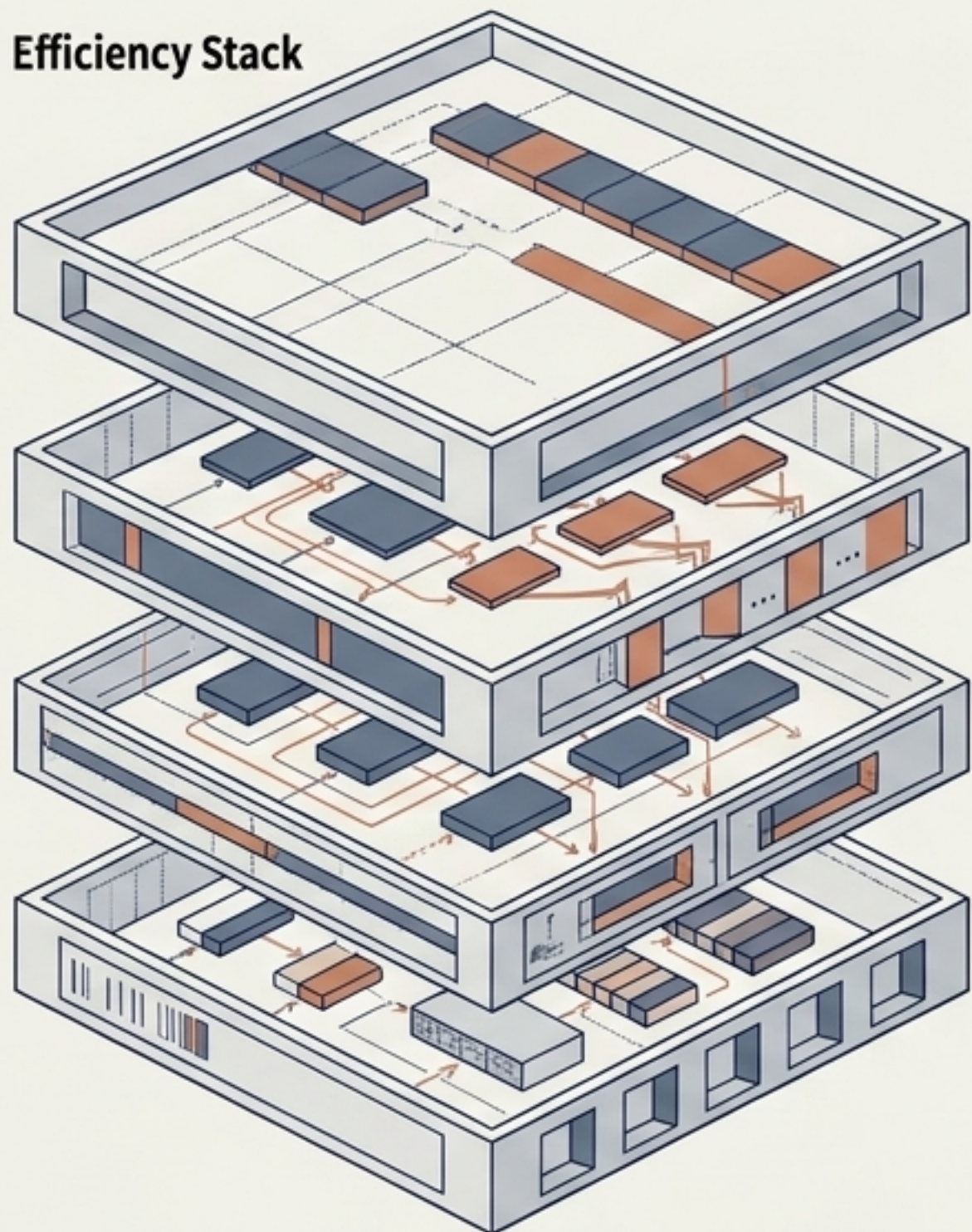
市場動向：OpenRouter上のオープンソース・トークン比率は、2026年1月の34%から6月には65%へ急増。人気上位4モデルを中国系が独占。

「20分の1の価格」を支える推論アーキテクチャ

中国モデルの低価格はダンピングではなく、「推論経済性そのものを設計目標」とした多層的な技術最適化の成果である。

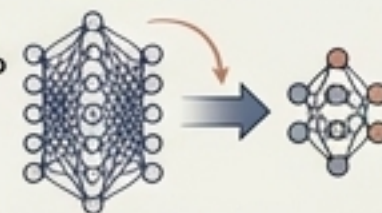
コストと資源要求の劇的な削減

Efficiency Stack



Layer 4: 知識蒸留 (Knowledge Distillation)

高性能教師モデルからの能力抽出。
高性能と軽量化の両立。



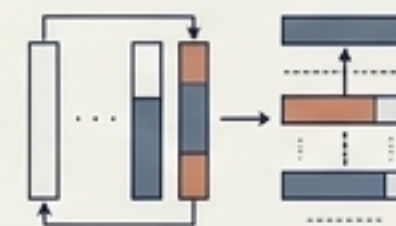
Layer 3: MoE (Mixture of Experts)

巨大な総パラメータに対し、実行時（活性）パラメータを極小化。
(例: DeepSeek-V4-Proは1.6T総パラメータに対し、わずか49Bのみ活性化)



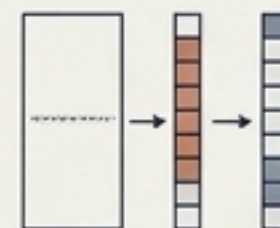
Layer 2: 疎注意機構・MLA (Memory Reduction)

KVキャッシュを圧縮・分離。
長文脈時のメモリ枯渇を防止。
(例: Kimi K2.7 Mooncake, GLM-5.2 IndexShareは1M文脈時の per-token FLOPsを2.9倍削減)



Layer 1: 量子化 (Quantization)

FP4/FP8混合精度の採用 (AWQ等)。
帯域と格納コストを劇的に削減。
(例: DeepSeek-V4)



結論:
能力の上限は大きく、実行時負荷は小さく。「長く・多段で・繰り返す」エージェント推論に最適化された構造。

主要モデルの経済性とデプロイメント比較

モデル	Claude Fable 5 (米国 / Anthropic)	GLM-5.2 (中国 / Zhipu)	Kimi K2.7 Code (中国 / Moonshot)	DeepSeek-V4-Pro (中国 / DeepSeek)
コスト	\$10.00 / \$50.00	\$1.40 / \$4.40	\$0.95 / \$4.00	\$0.435 / \$0.87
技術仕様	1M文脈, 汎用フロンティア	1M文脈, IndexShare最適化	MoE (1T総/32B活性), MLA	MoE (1.6T総/49B活性)
デプロイメント	クローズド (APIのみ)	MITオープン (自社運用可)	Modified MIT (OpenAI互換)	MIT (vLLM等で自社運用可)

インパクト試算：月間100M入力・20M出力の社内検索タスクでの比較：
Claude Fable 5 (約\$2,000) vs DeepSeek-V4-Pro (約\$60.9) = **タスク完了あたりコストで約32倍の差。**

運用形態のジレンマと「出口の広さ」



API利用

- メリット: 立ち上がりが速い、最新モデルへの追従が容易。
- デメリット: 可用性リスク（停止）、価格改定の波及、越境移転によるデータ機密性リスク。

出口の広さ (Broad Exit Options)



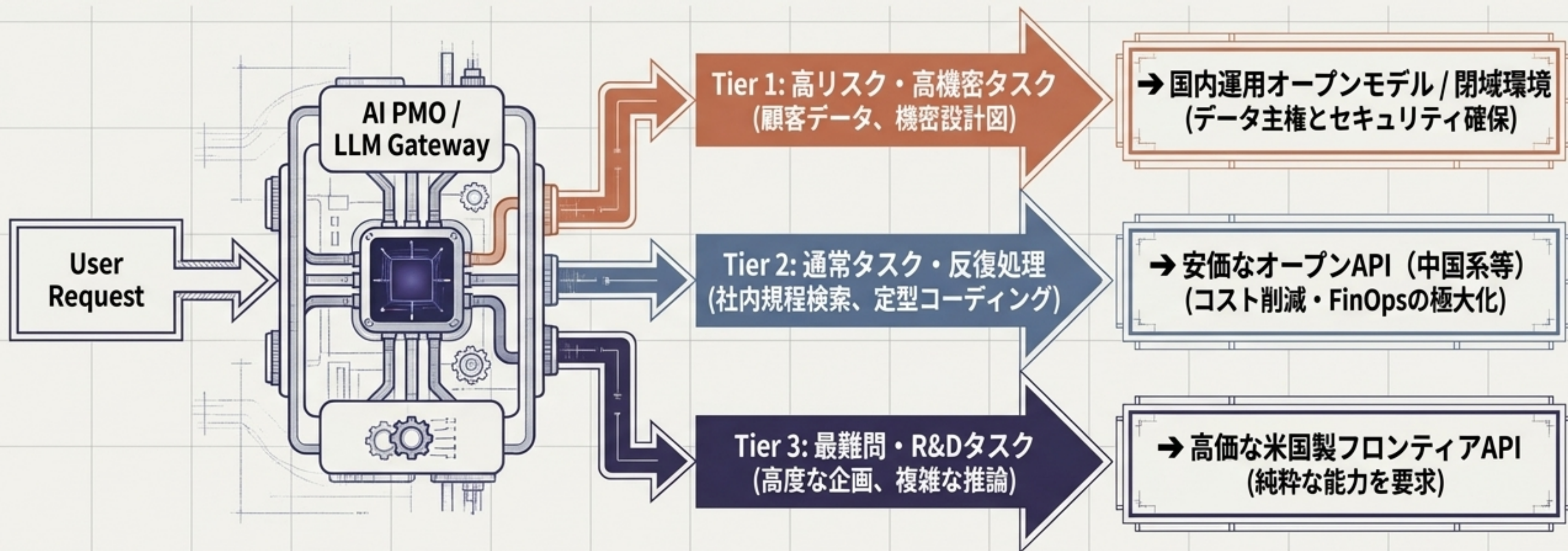
自社環境運用 (国内クラウド/オンプレ)

- メリット: モデル停止リスクの排除、越境移転リスクの回避、データ完全主権。
- デメリット: GPU・推論スタック・SRE等の高い固定インフラ費。

オープンモデル最大の強みは「出口の広さ」。最初は手軽なAPIで検証・評価を行い、本番稼働や高機密データ適用時には国内環境（自社運用）へ移行できる。閉鎖APIには不可能な調達オプション。

エンタープライズ・アーキテクチャ：3層ハイブリッドルーティング

米CoinbaseはLLMゲートウェイのデフォルトをオープンモデル（GLM/Kimi等）に切り替え、AI支出を半減させた。日本企業も用途別の最適ルーティングが急務である。



企業は直ちに用途、機密性、コスト上限、監査ログ等の「全社共通モデル選定基準」を策定すべきである。

業種別インパクト：汎用チャットから「現場の実務」へ

製造 (Manufacturing)



最重要領域（日本の勝ち筋）

現場データのAI-Ready化とフィジカルAIへの投資。汎用AIではなく、工場・自動化・品質・保全・工程最適化で勝負する。

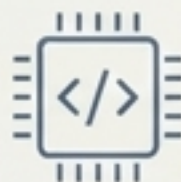
金融 (Finance)



ハイブリッド運用

高規制の顧客説明や審査は「人間承認+閉域最先端」、社内コードや規程検索は「オープンモデル」で切り分ける（金融庁方針に準拠）。

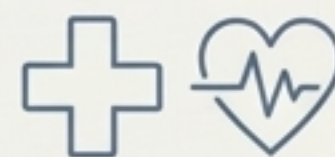
IT・ソフトウェア (IT & Software)



直接的なコスト革命

長文脈・コーディング特化の廉価モデル（Kimi/DeepSeek等）へ通常タスクを流し、人件費あたりの開発速度とROIを最大化する。

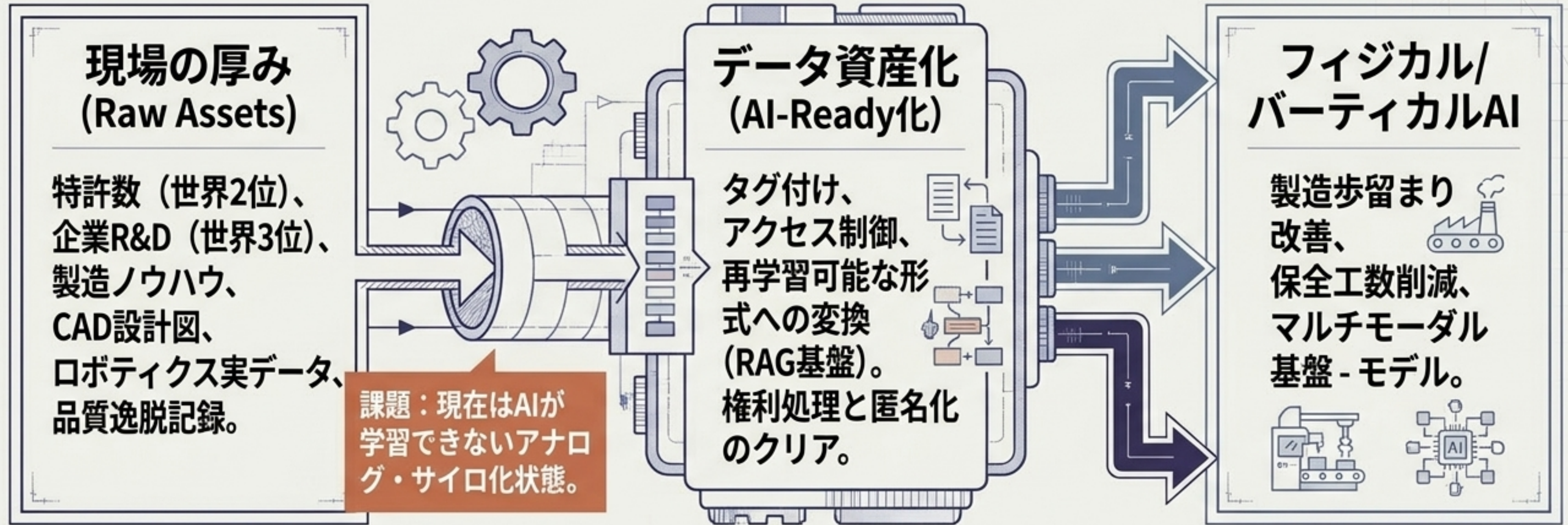
ヘルスケア (Healthcare)



厳格な検証と用途限定

臨床判断（自律診断）への即時採用は回避。まずは薬事・治験文書、学会要約、品質文書などの「業務生産性向上」に特化する。

日本の固有資産：「汎用AI」ではなく「AI-Ready化」で勝つ



日本の優先順位は「モデルを増やすこと」ではなく、
「現場の暗黙知をAIが食べられる形（データ循環ループ）にすること」である。

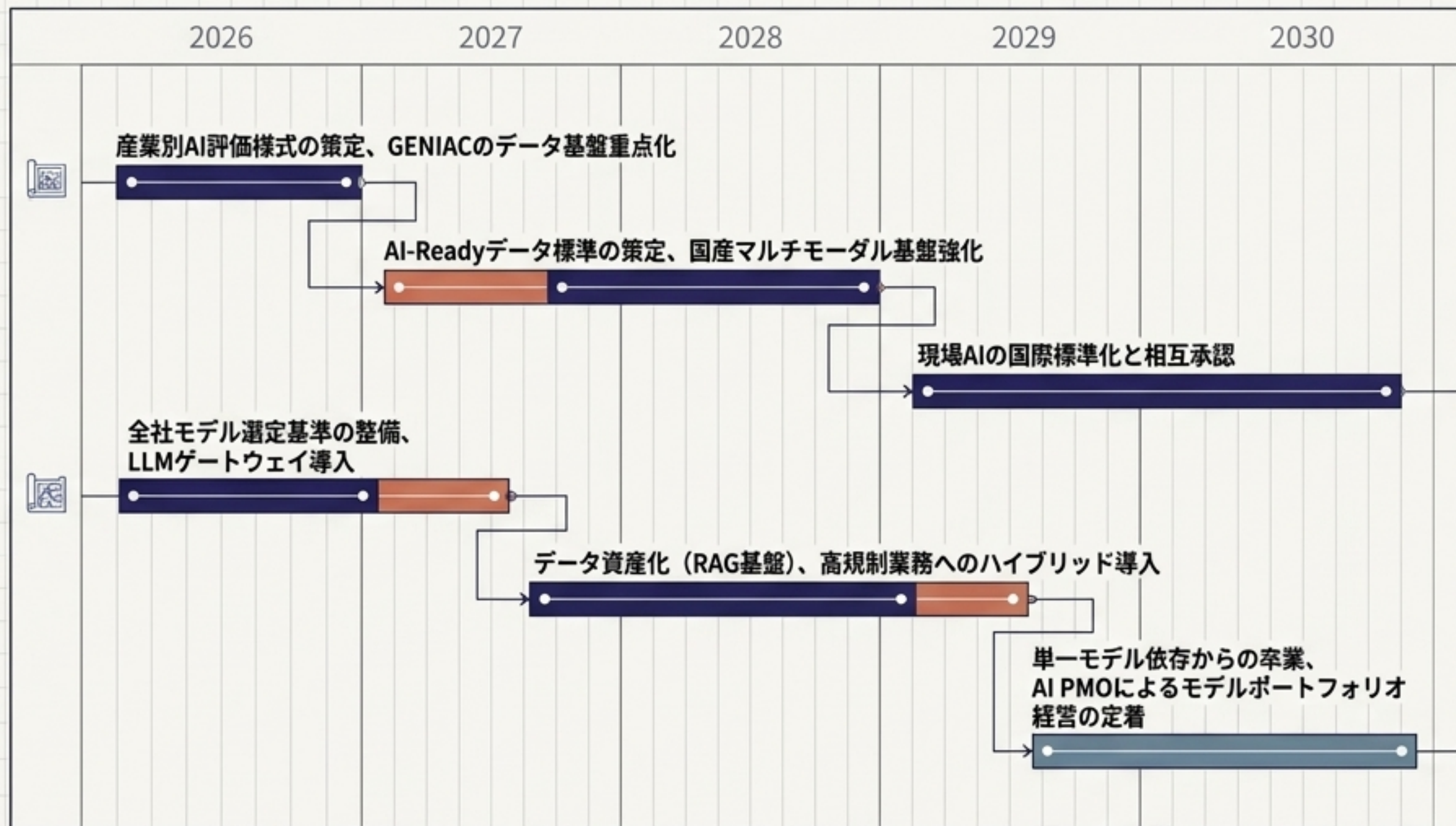
セキュリティ・規制・サプライチェーンの統制

評価標準と契約実務の整備は、モデル開発競争と同格の経営課題である。

リスク分類	影響度	企業側の対策
海外APIの停止・制限（可用性）	● [影響：高]	マルチモデル化、代替モデルの常備、自社運用先の確保。
個人データの越境移転（主権）	● [影響：高]	APPI（個人情報保護法）準拠の国内運用、匿名化、DPA/契約整備。
サプライチェーン偏在（ハードウェア）	● [影響：高]	GPU不足に備えた推論アーキテクチャの工夫、国内外の複線化。
モデル由来・知財紛争（正統性）	● [影響：中～高]	知識蒸留リスクの回避、ベンダー審査、補償条項の確認。
機密コード・情報の漏洩（セキュリティ）	● [影響：高]	社内閉域、RAG境界制御、権限分離。

EU型の事前許認可でも米国型の市場任せでもない、
業界別の実務型ガバナンスが日本の武器となる。

実行ロードマップ：政府と企業の協調（2026-2030）



短期:

- [政府] 産業別AI評価様式の策定、GENIACのデータ基盤重点化。
- [企業] 全社モデル選定基準の整備、LLMゲートウェイ導入。

中期:

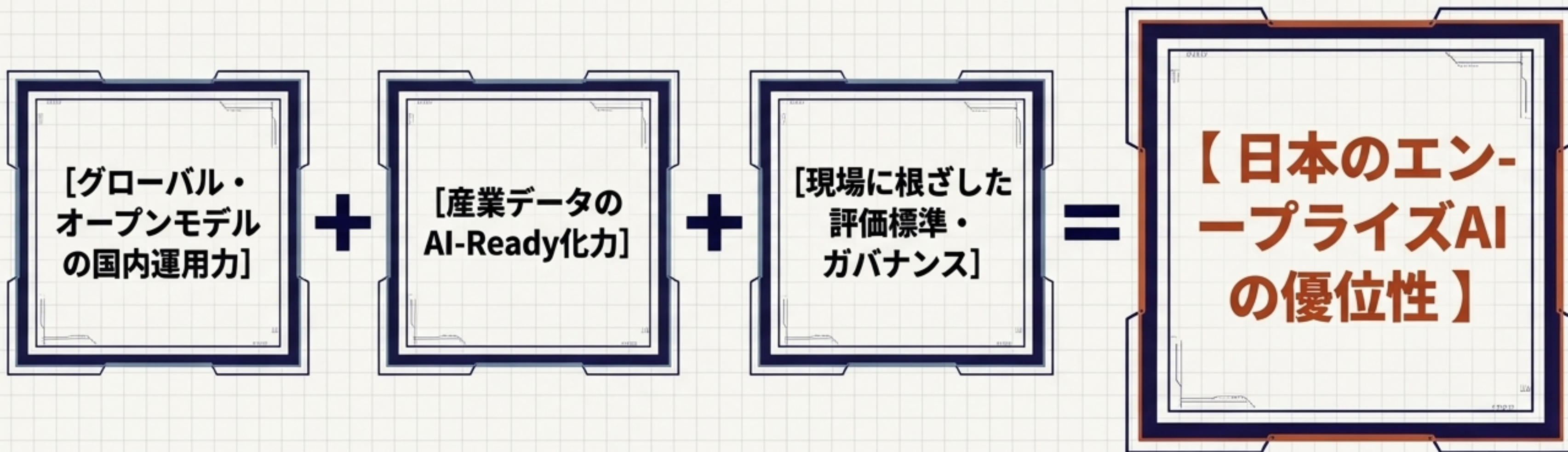
- [政府] AI-Readyデータ標準の策定、国産マルチモーダル基盤強化。
- [企業] データ資産化（RAG基盤）、高規制業務へのハイブリッド導入。

長期:

- [政府] 現場AIの国際標準化と相互承認。
- [企業] 単一モデル依存からの卒業、AI PMOによるモデルポートフォリオ経営の定着。

結論：日本が描くべき「勝利の方程式」

日本は、米国型の「閉鎖フロンティア依存」にも、
中国型の「低価格オープン依存」にも寄り切るべきではない。



足りないのは技術そのものではない。すでにGENIACや現場の特許・知見という土台はある。
今求められているのは、経営層の「動員の速さ」と「実装の一貫性」である。