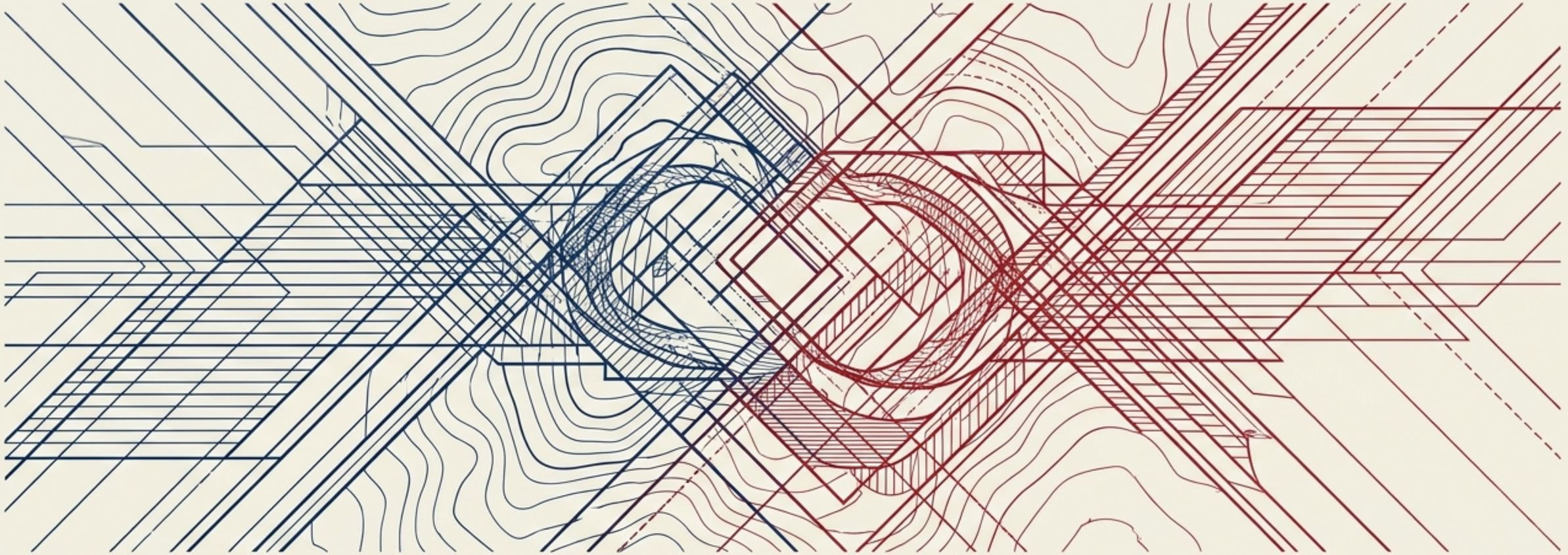
米中フロンティアモデル競争の現在地と次世代ブレイクスルー予測

「半年遅れ」神話の終焉と、2026-2029年の非対称なAIパラダイムシフト



CONFIDENTIAL / INTERNAL BRIEFING

Date: 2026-08-16

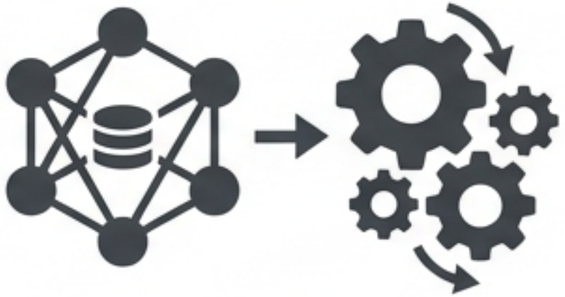
Executive dashboard

The Myth is Dead
神話の崩壊

「中国は米国から半年遅れ」という認識は、もはや過去の遺物である。

総合知能スコアにおける米中の差はわずか1〜3ポイントへ縮小。

米国最上位（Claude Opus 5）と中国最上位（Kimi K3）は実質的に同一世代の競争バンドに突入した。



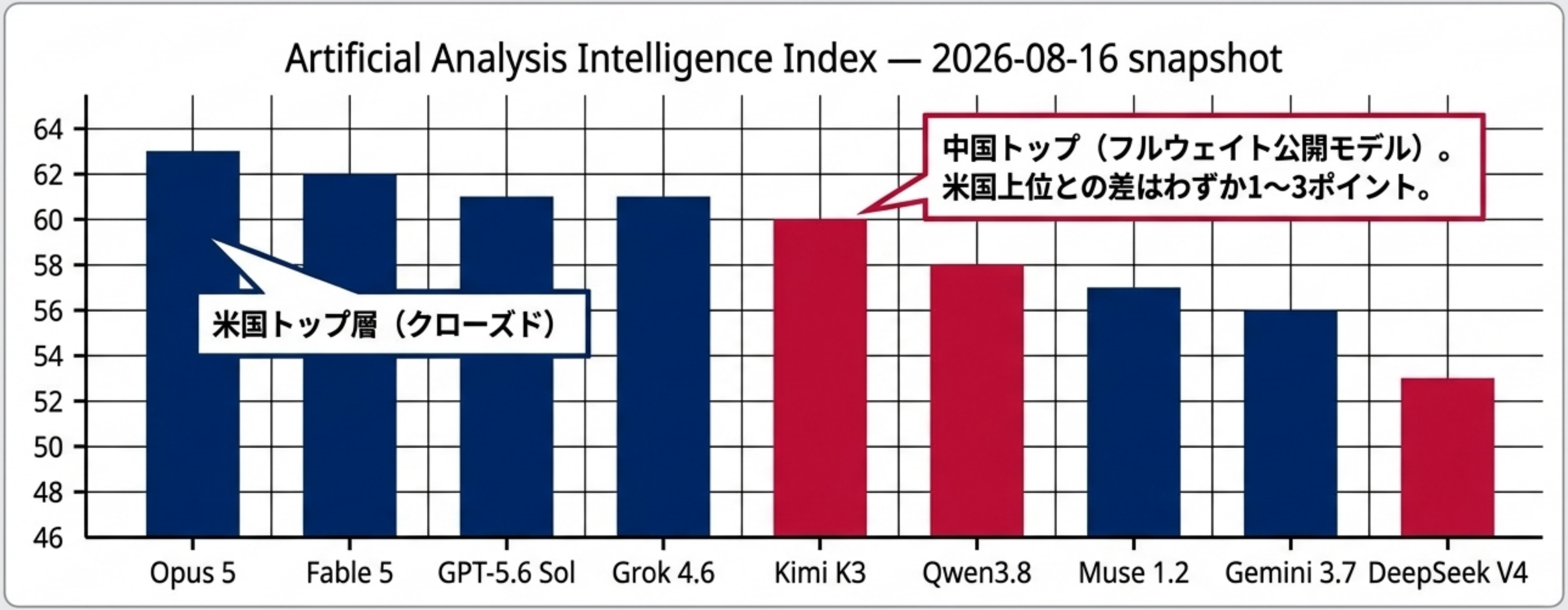
The Paradigm Shift
競争軸の変化

「巨大事前学習モデルの構築」から、推論時の計算量（Test-time compute）の動的配分と、数日稼働するエージェント能力へと根本的に移行した。

The Asymmetric Warfare
非対称な戦い

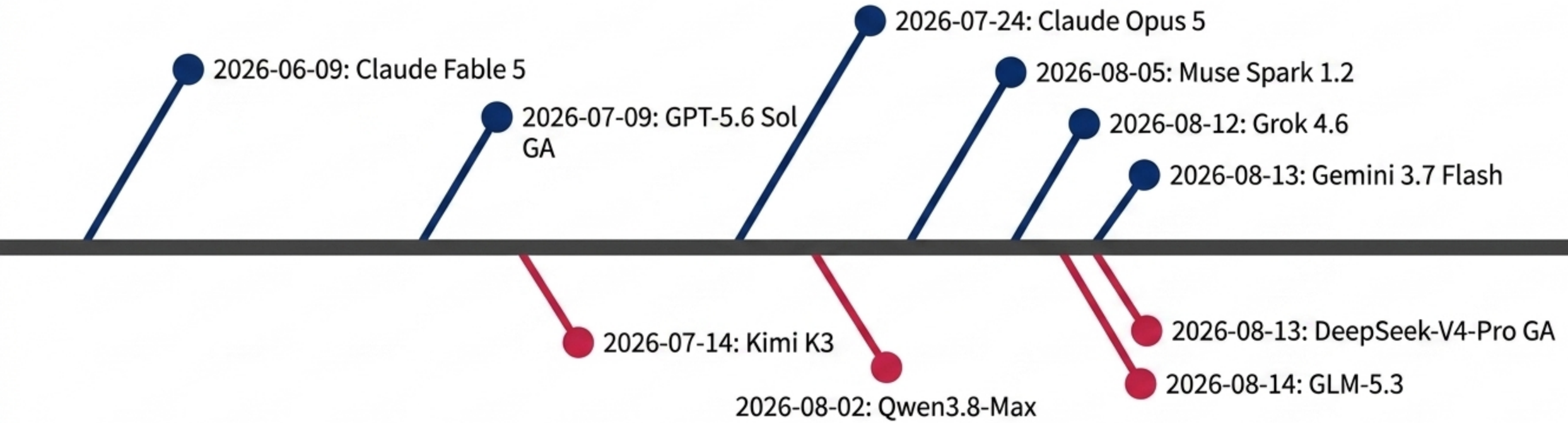
米国は「最先端ハードウェアとフルスタック統合」で優位に立ち、中国は「オープンウェイトとアルゴリズムによる極限の計算効率化」で相殺する非対称戦となっている。

知能指数の現在地：同一世代の競争バンドへ突入



2025年までの「米国クローズドが先行し、数ヶ月後に中国オープンが追う」構造は崩壊。現在、オープンウェイトのKimi K3（60）が米国最上位のクローズドモデルと同等の知能水準で競争している。

リリース密度の真実：2026年夏の同時多発的フロンティア競争



中国4モデルは約1ヶ月（31日間）に集中、米国も2ヶ月以内に6モデルをリリース。「半年の固定ラグ」は観測されず、互角のリリース速度（Cadence）を示している。

非対称な戦力図：各陣営の「勝ち筋」の違い

評価軸	米国優位	互角	中国優位	根拠データ
総合モデル能力	●			AA top 63（米）vs 60（中）
コーディング / 長期エージェント		●		Kimi 88.3 vs 米国各種トップ層。首位が頻繁に入れ替わる
商用推論速度 / フルスタック	●			Gemini Flash 340 tok/s, GPT-5.6 750 tok/s vs 中国勢 40-50 tok/s
先端ハードウェア基盤	●			NVIDIA Rubin 2026 H2 vs Huawei Ascendの制約環境
推論効率化アルゴリズム			●	Kimi MXFP4 QAT, DeepSeek DSA, GLM IndexShare
オープンウェイト・フロンティア			●	Kimi full weights, DeepSeek MIT vs 米国は全クローズド
安全性評価の透明性	●			米国固有System Card公開 vs 中国Trusted Access段階公開

米国フロンティアモデル・スペックマトリクス（2026.08）

モデル	パラメータ数 / コンテキスト	主要技術・機能	商用形態
GPT-5.6 Sol	未公開 / 1.05M ctx	Reasoning effort (none〜max), Ultra subagents	Closed (API/ChatGPT)
Claude Opus 5	未公開 / 1M ctx	Adaptive reasoning, Test-time compute拡大	Closed ($5/$25 MTok)
Claude Fable 5	未公開 / 1M ctx	高リスクドメインのFallback機構	Closed (Premium tier)
Gemini 3.7 Flash	未公開 / 1M ctx	Native multimodal, 高スループット重視	Closed ($0.75/$3.75 MTok)
Grok 4.6	未公開 / 500K ctx	Agentic RL, 再生成SFT trajectories	Closed ($2/$6 MTok)
Muse Spark 1.2	未公開 / 1.049M ctx	Coding workflow強化, Muse Code連携	Closed

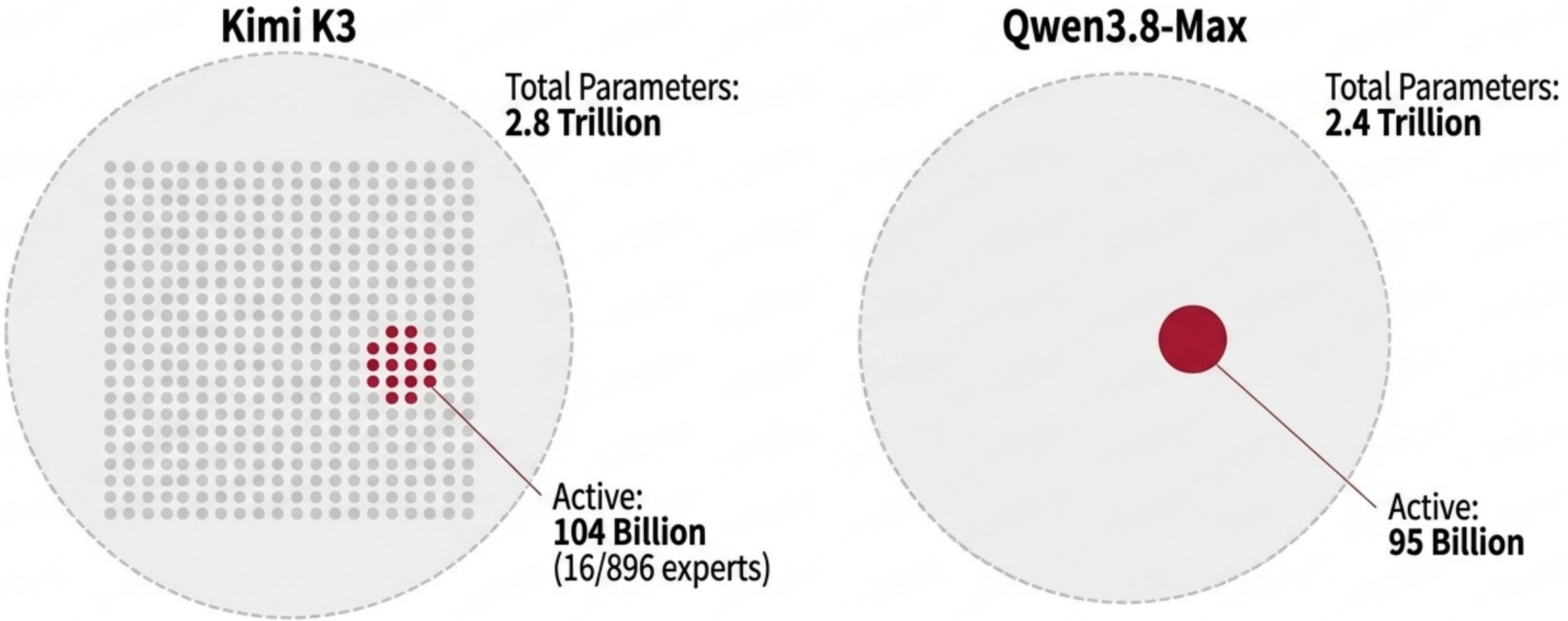
米国モデルは意図的にパラメータ数を非公開化（Closed Proprietary）。推論速度とフルスタック統合（Cerebras等）で圧倒的なAPIパフォーマンスを誇る。

中国フロンティアモデル・スペックマトリクス（2026.08）

モデル	アーキテクチャ (Total / Active)	推論最適化技術	商用形態・ライセンス
Kimi K3	2.8T Total / 104B Active (16/896 experts)	MXFP4/8 QAT, Kimi Delta Attention	Full weights公開 (Custom License)
Qwen3.8-Max	2.4T Total / 95B Active	FP8対応, Massive sparse MoE	2.4T-A95B weights公開
DeepSeek-V4-Pro	1.6T Total / 49B Active	DeepSeek Sparse Attention (DSA)	Open weights (MIT License)
GLM-5.3	744B Total / 40B Active (GLM-5系)	IndexShare, MTP	Weightsは安全評価後公開予定

中国モデルの最大の特徴は「極端なSparsity（スパース化）」。総パラメータ数は巨大だが、推論時に稼働するActiveパラメータを極限まで絞り込んでいる。

中国の生存戦略：「巨大化」と「軽量化」のパラドックス



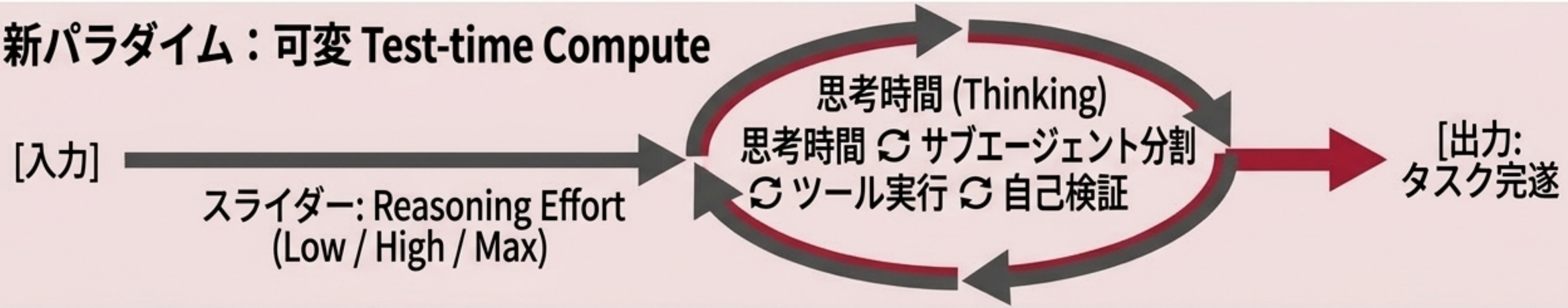
先端GPUへのアクセス制約（輸出規制）に対する中国の戦略的回答。モデルを極限まで巨大化しつつ、1トークンあたりの実計算量（Active FLOPs）を低精度化（MXFP4）やSparse Attentionで極限まで抑え込む。「ハードウェアの不利をアルゴリズムで相殺する戦略」。

新たな競争軸：静的推論から「動的Test-time Compute」へ

旧パラダイム：静的推論（Static Forward Pass）

[入力: タスクの難易度を問わず固定] → [巨大モデルで1回だけ推論] → [出力]

新パラダイム：可変 Test-time Compute



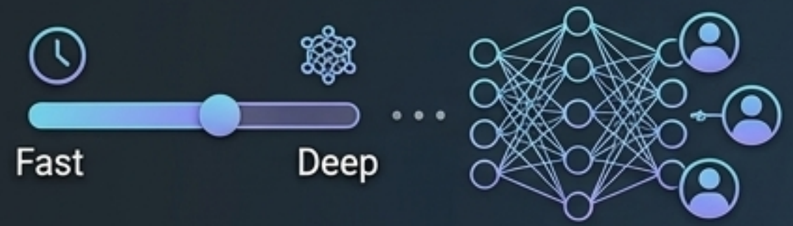
GPT-5.6 Sol: Max / Ultra (subagents追加)	Claude Opus 5: Adaptive reasoning	Kimi K3: Max thinking	DeepSeek V4: Low / High / Max

一つの固定モデルを一回フォワードする形から、タスク価値に応じて推論リソースを動的に配分する「AI OS的構造」へ収斂している。

次世代モデルを駆動する「4つの技術ドライバー」

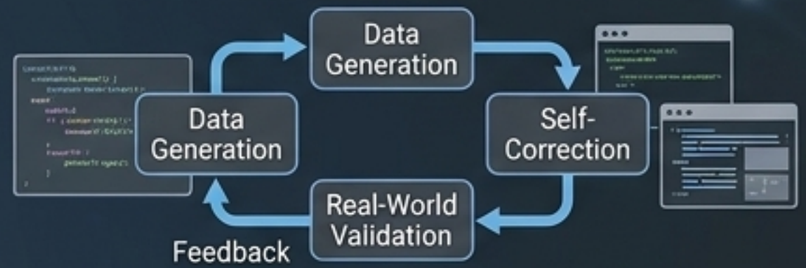
1. Test-time Computeの標準化

- 単一モデルでFast〜Deep〜Multi-agentを連続制御。
  推論時の計算量（Compute Scaling）が最大の競争軸に。



2. 合成データと実世界検証（Synthetic & Verifiable Data）

- Webテキストの枯渇を超え、モデル自身が推論軌跡を生成。
  実世界ワークフロー（Code/GUI）の自己採点ループ。



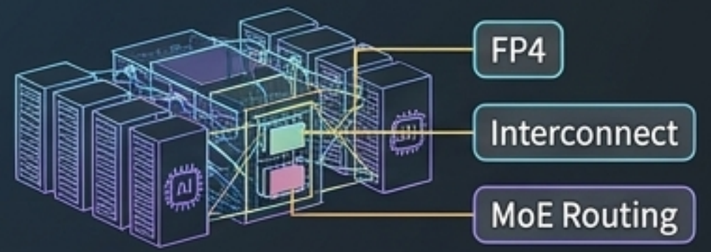
Next-Gen AI

3. AttentionとMemoryの構造改革

- 1Mコンテキストの力技から、必要な過去情報だけを選択
  ・圧縮する永続的メモリへ。KVキャッシュの極限圧縮。

4. HW-SWの垂直統合(Co-design)

- 単なるGPU性能ではなく、FP4、インターコネクト、MoE
  ルーティングを含むAIファクトリーシステム全体の競争。



ブレイクスルー・シナリオ I（短期：2026 H2 – 2027 H1）

2026 H2

2027 H1

確度：高

Dynamic Reasoningの標準化

ベンチマークの素点コスト増大のユーザー受容性が鍵。

確度：高

Dynamic Reasoningの標準化

ベンチマークの素点より「Task completion / $」が競争軸に。推論コスト増大のユーザー受容性が鍵。

確度：高

自律的コーディングエージェントの実用化

数時間〜数日稼働するエージェント (Kimi 48時間チップ設計、Grok長期RL)。自己テストやロールバックを標準搭載。

確度：高

FP4と巨大Sparse MoEの主流化

中国はM曖昧に採用。ハードウェア間の移植性がリスク。

確度：高

FP4と巨大Sparse MoEの主流化

中国はGPU制約緩和のため、米国はRubin規模での増幅のために採用。ハードウェア間の移植性がリスク。

確度：高中

Persistent Memoryへの移行

単なる長文脈（1M context）処理から、巨大な社内リポジトリやDBを選択・再利用する永続的メモリへ進化。

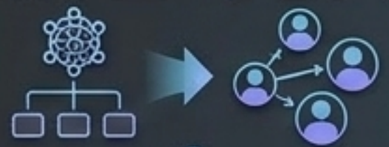
ブレイクスルー・シナリオ II（中長期：2027 H2 – 2029）

2027 H2

2029

確度：中高

階層型 マルチエージェント



確度：中高

階層型マルチ エージェント

Planner, Coder, Critic等による自律的オーケストレーション。企業業務のエンドツーエンド自動化。

確度：中高

AI主導のR&D データフライホイール

AIがコンパイラやベンチマークを自作し、次世代モデルを自ら改善。中国が計算リソースの不利を相殺する可能性。

確度：高

Cyber能力の アクセス制御

確度：高

Cyber能力の 閾値とアクセス制御

能力向上に伴い、「研究版」「Trusted版」「Public版」の分離が一般化。政府規制の介入ラインが焦点。

確度：低中

Transformer+ MoEからの脱却

KDA/DSAなどフルアテンションからの離脱が加速。計算量増加のルールが変われば、ハードウェアギャップがリセットされる可能性。

評価指標と経済性のシフト：「点数」から「業務の完遂コスト」へ

旧来の評価 (Static QA)

指標：$/MTok (入力トークン単価)

ベンチマーク：MMLU, HELM

マインドセット：「安価なモデルで100回ツールを呼び出して試行錯誤する」

新たな評価 (Real-world Workflow)

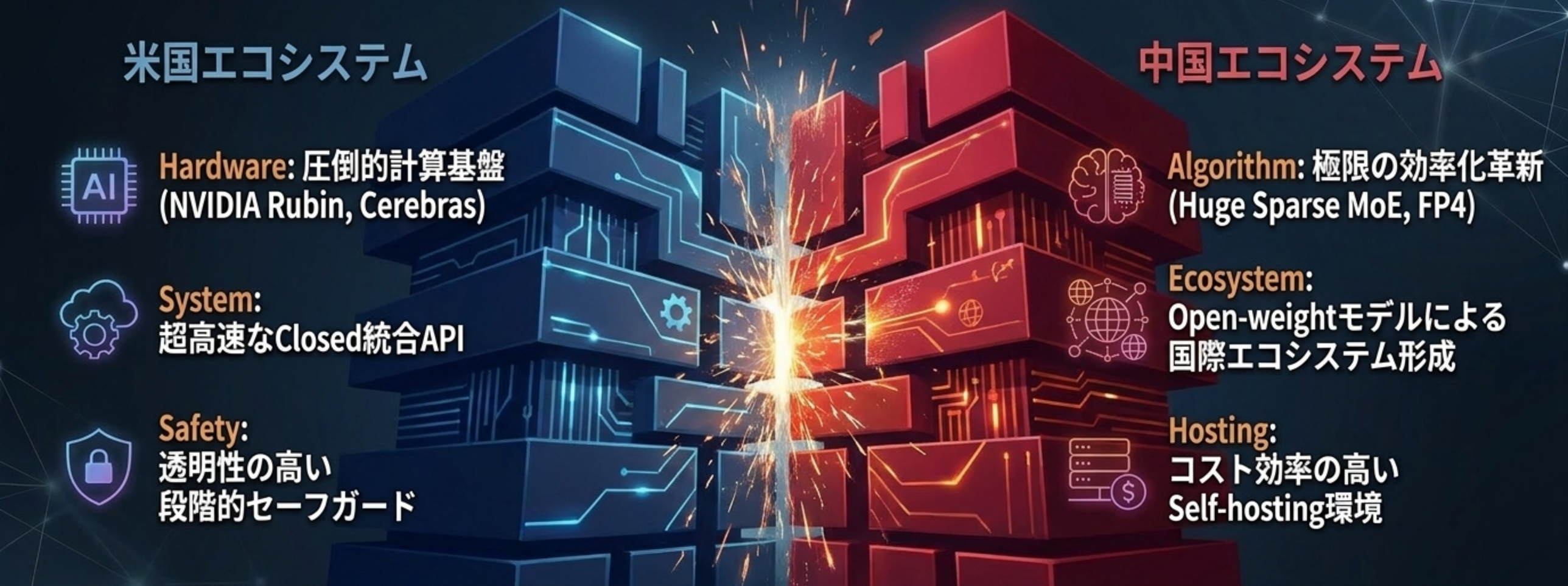
指標：$/successful-task, Human minutes saved

ベンチマーク：Terminal-Bench, OSWorld, GDPVal

マインドセット：「単価が2倍でも、一度で正しくタスクを完遂し再試行をゼロにするモデルの方が圧倒的に高コスパ」

モデル知能の進化により、ビジネス価値の源泉は「推論APIの安さ」から「エラーリカバリを含めた自律的タスク完遂能力」へと完全に移行した。

総合結論：2027年、「AIスタック圏」の非対称な激突



米国エコシステム

Hardware: 圧倒的計算基盤
(NVIDIA Rubin, Cerebras)

System:
超高速なClosed統合API

Safety:
透明性の高い
段階的セーフガード

中国エコシステム

Algorithm: 極限の効率化革新
(Huge Sparse MoE, FP4)

Ecosystem:
Open-weightモデルによる
国際エコシステム形成

Hosting:
コスト効率の高い
Self-hosting環境

単一の勝者は存在しない。米国はハードウェアの力で押し切り、中国はアルゴリズム効率で相殺する。2027年のAI市場は、アーキテクチャ、エージェント、コスト、安全性の各軸で首位が頻繁に入れ替わる**「多極的かつ非対称なAI圏競争」**となる。