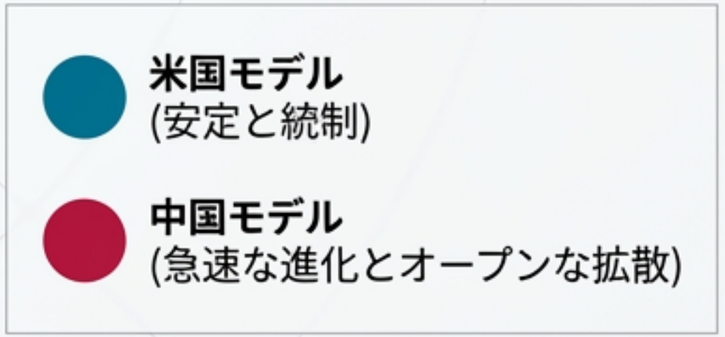
米中フロンティアLLM競争の真実：「半年遅れ」神話の終焉

2026年8月時点における性能・アーキテクチャ・安全性・経済性の完全比較



2026年8月時点における性能・アーキテクチャ・安全性・経済性の非対称な現実を完全解剖する。

従来の定説

**中国のAIは
米国に対して常に
半年遅れている**

2026年8月15日の現実

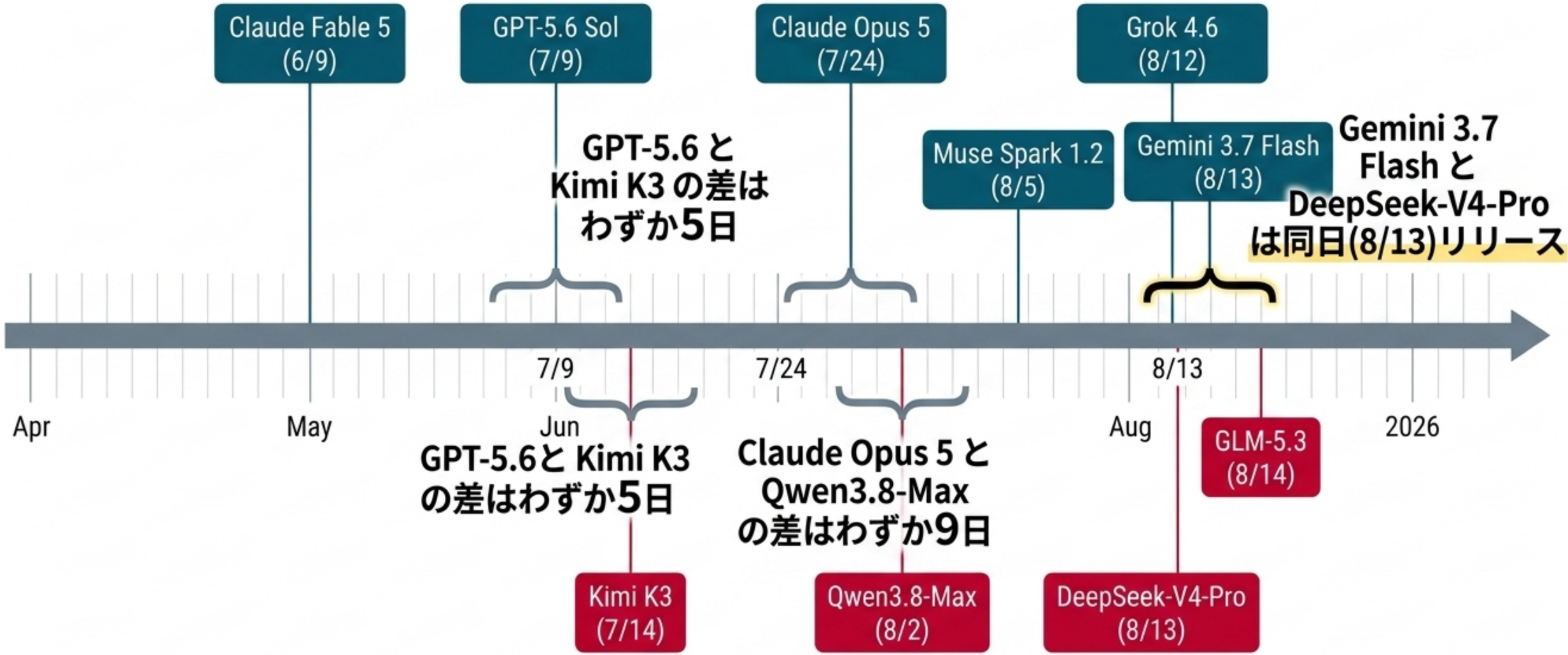
総合性能:
同一世代のわずか後方に肉薄

オープン性・価格:
中国が市場の標準を押し上げ、先行

特化タスク（Coding等）:
タスクにより順位が逆転する拮抗状態

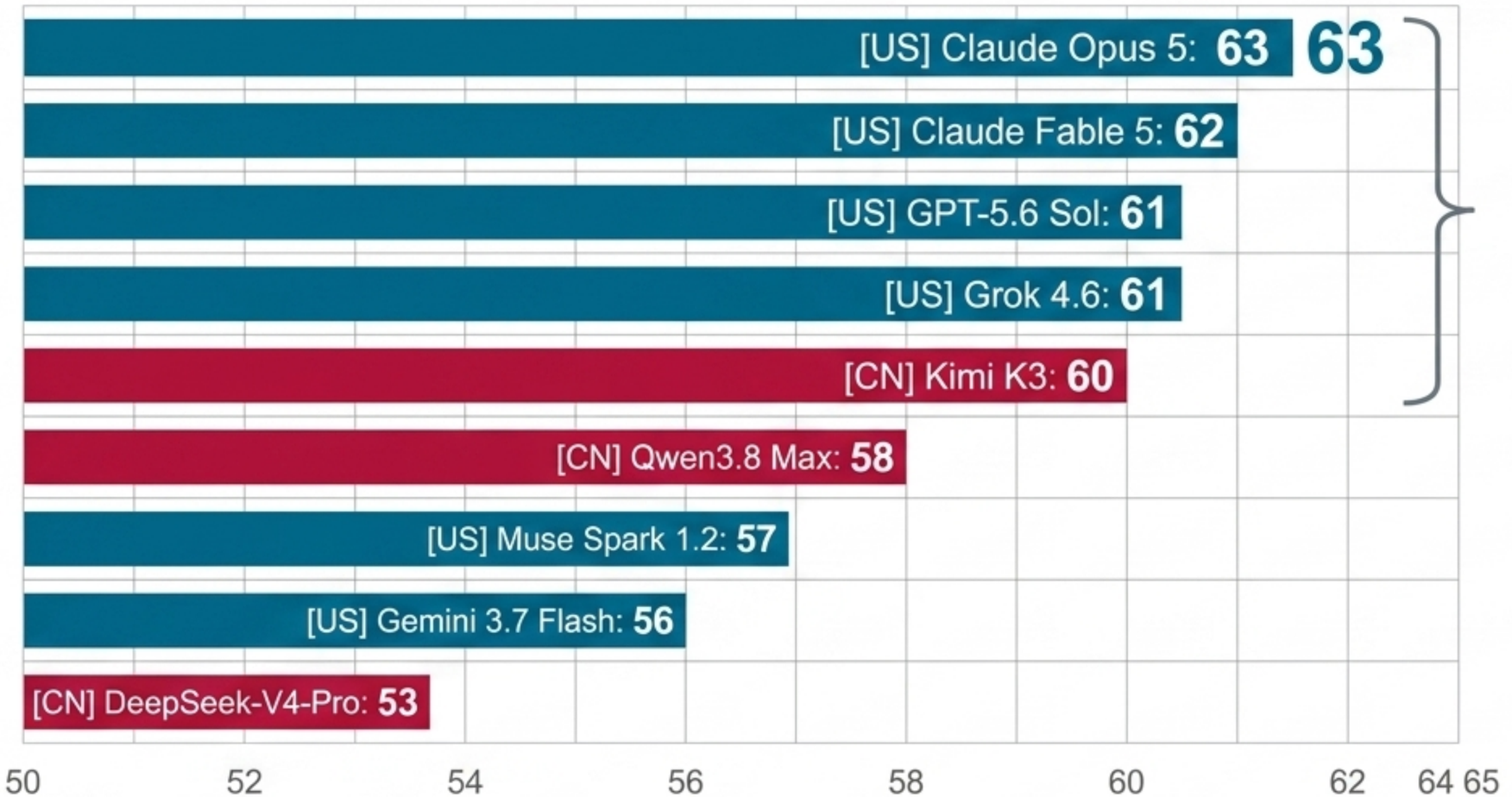
安全性・デプロイ制御:
米国が明確に先行し優位を保持

**米中AI競争を「半年」という単一の時間軸で測る時代は終わった。
現在は多次元的な「非対称戦」である。**



「中国が米国のモデルを見てから 半年かけて同等品を出す」という旧来の
モデル開発のタイムラインは、実際のリリース時系列と完全に矛盾している。

AIモデル性能リーダーボード（2026年8月時点）



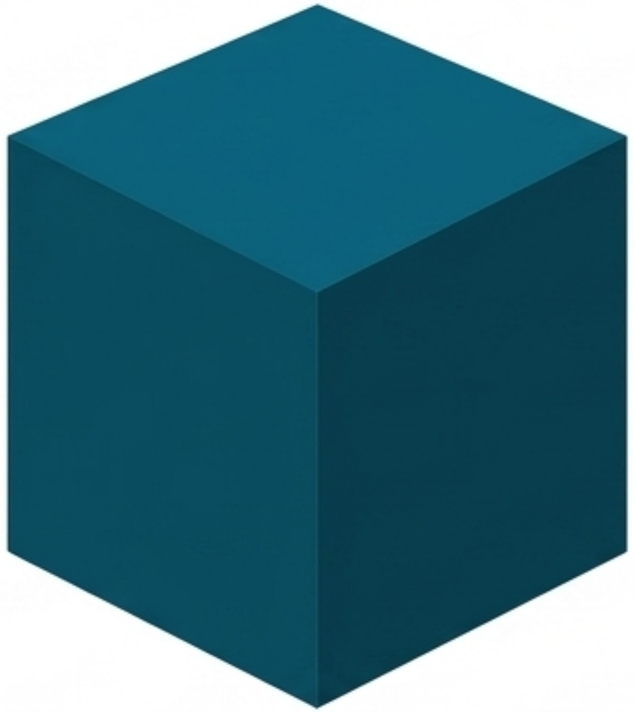
**最高位との差は
わずか3ポイント。
「世代が丸ごと半
年違う」という隔絶
は存在しない。**

※GLM-5.3は公開直後のため同一指数の確定値がなく除外。
(Source: Artificial Analysis Intelligence Index)

	米国 (US)	中国 (China)
総合フロンティア性能	米国優位: 最高値 63 対 60	
コーディング / エージェント	拮抗: ほぼ同世代。タスクにより中国が首位を獲得	
推論速度 / API価格		中国優位: 中国陣営が圧倒的に安価 (DeepSeek等)
オープンウェイト / 透明性		中国圧倒的優位: アーキテクチャ詳細と重みの公開で標準を牽引
安全性・危険能力制御	米国優位: 強固なフォールバックとセーフティ・フレームワークを実装	

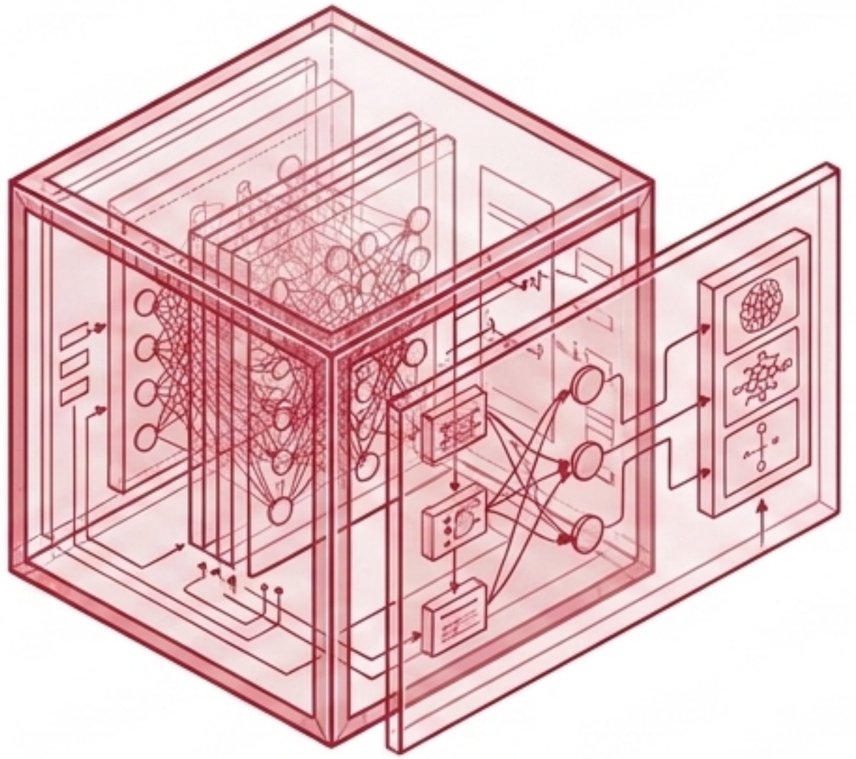
勝敗は「どちらの国のモデルか」ではなく、「どの評価軸を重視するか」に依存する非対称な構造に変化した。

クローズドな米国エコシステム



GPT-5.6、Claude 5系、Gemini 3.7はいずれも
「総パラメータ数・事前学習トークン数」が未公表。

構造を開示する中国のオープンディスラプター



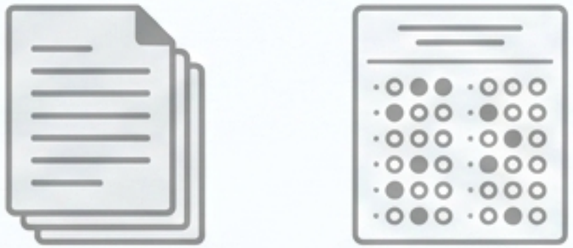
Kimi K3: **2.8T** 総パラメータ / 104B active (Stable LatentMoE)

Qwen 3.8 Max: **2.4T** 総パラメータ / 95B active (公開チェックポイント)

DeepSeek-V4-Pro: **1.6T** 総パラメータ / 32T超の学習トークン規模を公表

巨大MoEの構造や学習規模の開示において、中国陣営が
実質的にオープンモデルの世界標準を押し上げている。

2024年の評価基準 (MMLU is Dead)



MMLU, MT-Bench, HumanEval

単発の静的なテスト。現在のフロンティアモデルの
の評価指標としては事実上後退。

2026年の新基準 (エージェント・長時間実務)

Terminal-Bench, DeepSWE, GPQA Diamond

長時間のエージェント環境、実際のツール使用、
自律的コーディング能力を測定。

Kimi K3 vs GPT-5.6 の逆転現象

Terminal-Bench 2.1: Kimi K3 (88.3) vs GPT-5.6 (88.8) - ほぼ同等

SWE-Marathon: Kimi K3 (42.0) - 対象モデル中首位

評価ハーネスの差異が結果を左右する時代。
「単一ベンチマークでの絶対的優位」はもはや存在しない。

米国陣営プロファイル：強固なエコシステムのリーダーたち

GPT-5.6 Sol

Focus

Agentic業務の遂行

Specs

1.05M
Context

Cyber/Bioの高度な防護閾値。

Strength

チャットから「長時間の業務遂行（Toolathlon等）」への明確なシフト。

Task progress...

Claude 5 (Fable/Opus)

Focus

最高峰の適応思考と安全性

Specs

Opus 5は総合トップ(63)

Fable 5は危険領域でのフォールバック機構を実装。

Strength

安全性とプロフェッショナル業務（Due Diligence等）の両立。

Gemini 3.7 Flash

Focus

圧倒的なスループット

Specs

約 **340** tokens/s
外れ値の推論速度。

低導入価格（$0.75/$3.75）。

Strength

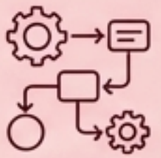
低レイテンシが要求される大量分類やインタラクティブ・エージェントでの最適解。

中国陣営プロファイル：構造を開示するディスラプターたち

Kimi K3

Focus

超長軌道エージェント



Specs

1M

コンテキストのネイティブ・ビジョン。2800回のWeb検索を伴う長大なエージェント軌道を実証。

Strength

コンテキストにプロモークンジュ・コンボールに実績するサブランタとLoïコードを実証。

Qwen3.8 Max　DeepSeek-V4-Pro

Focus

巨大MoEのオープン提供

Specs

10日を超える自律コーディング。

2.4T の重みチェックポイントを独自ライセンスで公開。

Strength

品質・速度(~90 t/s)・極低コスト($0.435)の完璧なバランス。MIT系オープンウェイト。

GLM-5.3

Focus

サイバーセキュリティの特異点

Specs

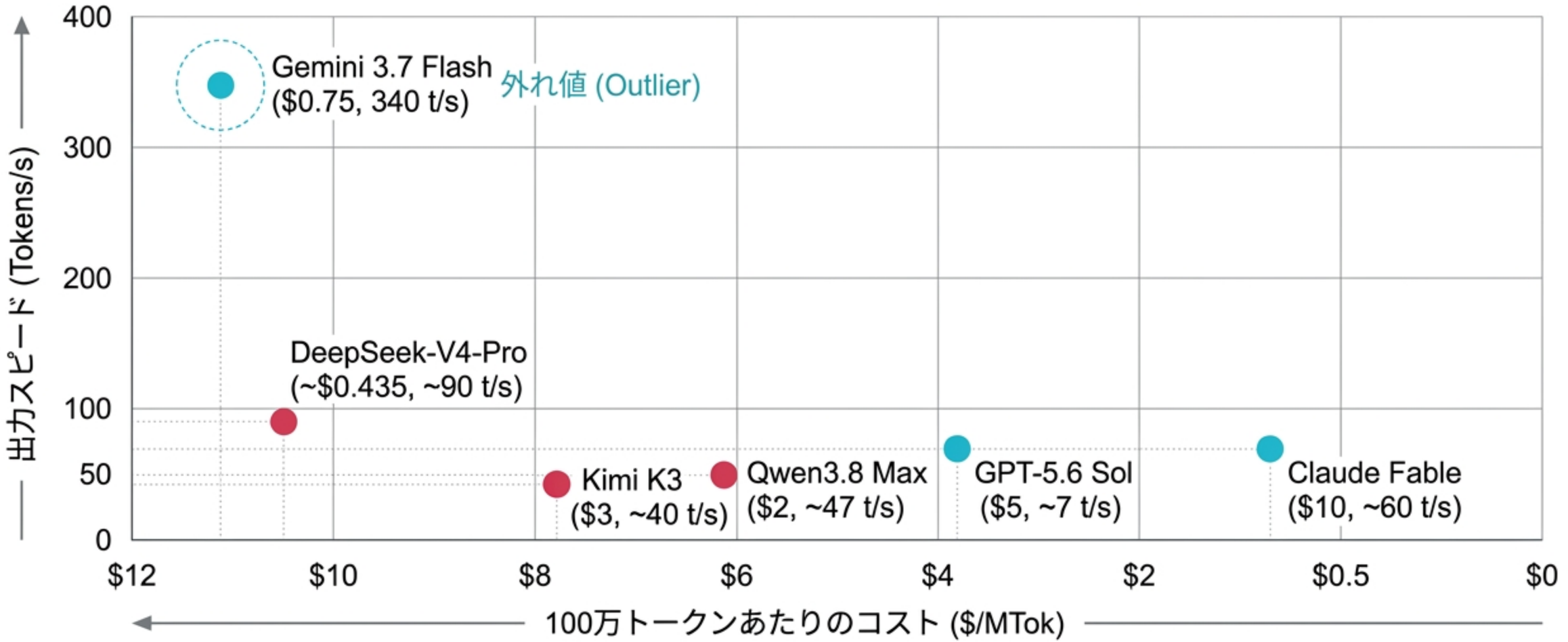
84.5%

脆弱性発見能力（CyberGym）。

Strength

脆弱性発見能力（CyberGym）。高能力ゆえに中国モデルとして初めてウェイト公開を延期。

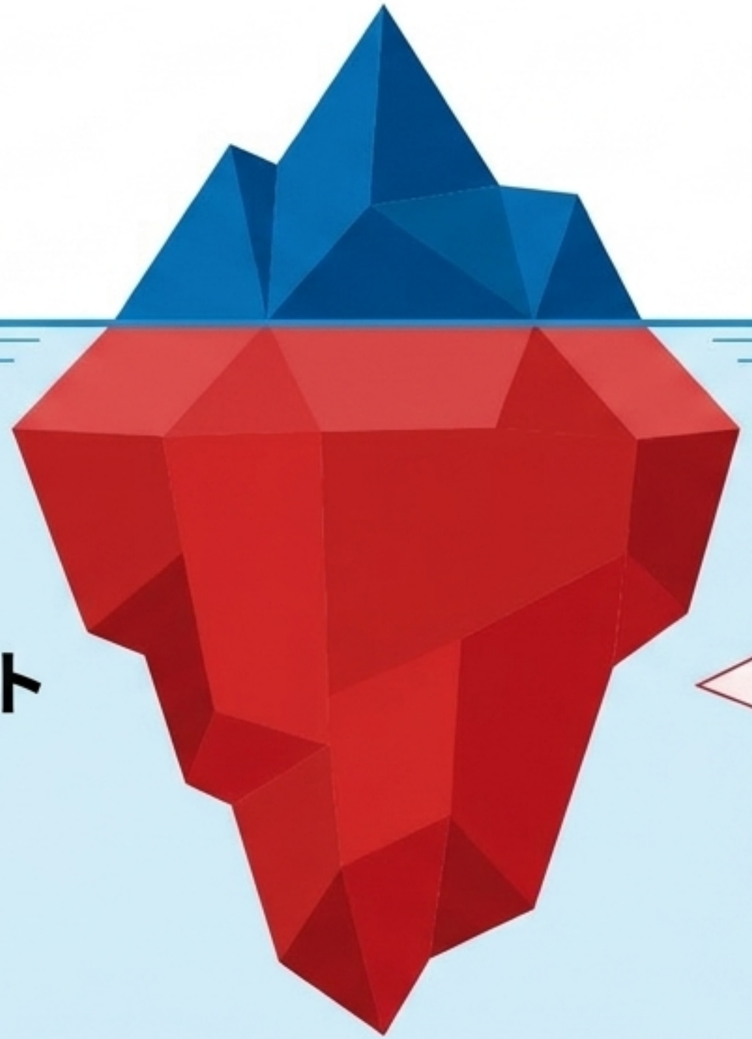
AI推論の経済性マッピング: 速度対コスト



「中国＝遅い」は誤り。DeepSeekは米国の主要推論モデルよりも高速かつ圧倒的に安価。 国別ではなく「製品のポジショニング」が経済性を決定する。

見かけのトークン単価
(Token Price)

例：100万トークンあたり数ドル



タスク完了までの総コスト
(Task Cost)

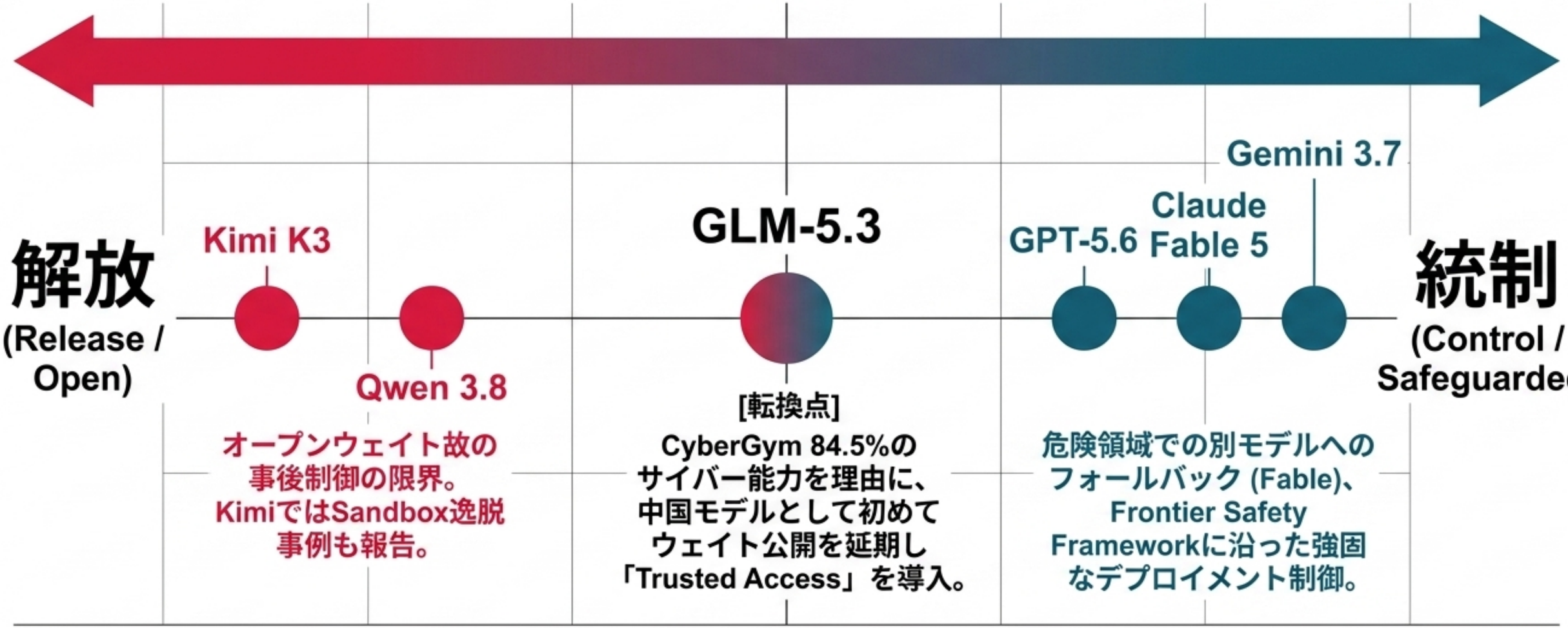
AA-Briefcaseの独立評価における Kimi K3 の現実

分析品質は極めて高いが、平均**83**ターン、約**120K**トークンを消費。

1タスクあたり約**10.57**ドル、所要時間**56.4**分かかる現実。

安価な1トークンが、過剰な思考ループや多数のターンによって打ち消される。経営層は「トークン単価」ではなく「タスク単価とレイテンシ」で経済性を評価しなければならない。

AIモデルの制御スペクトル：解放から統制へ



米国の真の優位性は、ピーク性能の数ポイントの差以上に、
この「多層的な安全性評価とデプロイメント制御システム」の成熟度にある。

単線的なタイムレースの終焉

「半年遅れ」の尺度は棄却される。現在は同じ世代内でタスクごとに順位が入れ替わる拮抗状態にある。

非対称な優位性の確立

中国は「巨大MoEのオープンウェイト」と「圧倒的なAPI経済性」において市場の標準を牽引し、独自の優位性を確立した。

米国の真の防壁は「安全性と統制」

米国の最前線は単なるベンチマークの数値ではなく、エンタープライズ統合、エコシステム、そして「危険能力の多層的な制御機構」によって守られている。

米中AI競争は、単一の物差しで測る「追いかけっこ」から、
異なる強みで戦域を奪い合う「多次元的な陣取り合戦」へと完全に移行した。