

米中フロンティアモデル比較調査：2026年 夏の最新ランドスケープ

単一の「半年遅れ」から、領域別の「**非一様な競争**」へのパラダイムシフト

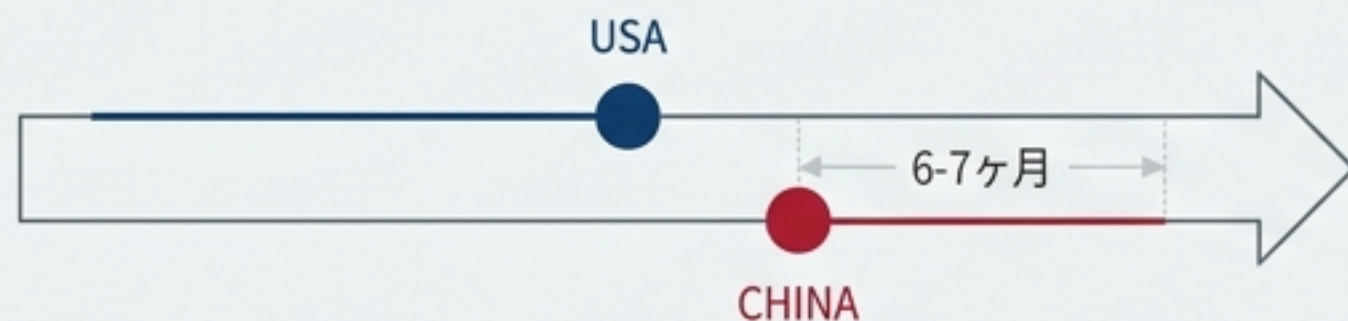
対象モデル: 米国上位6モデル (GPT-5.6 Sol, Claude Fable 5等) /
中国上位4モデル (Kimi K3, Qwen3.8-Max等)

評価基準日: 2026年8月15日

提供: Manus AI

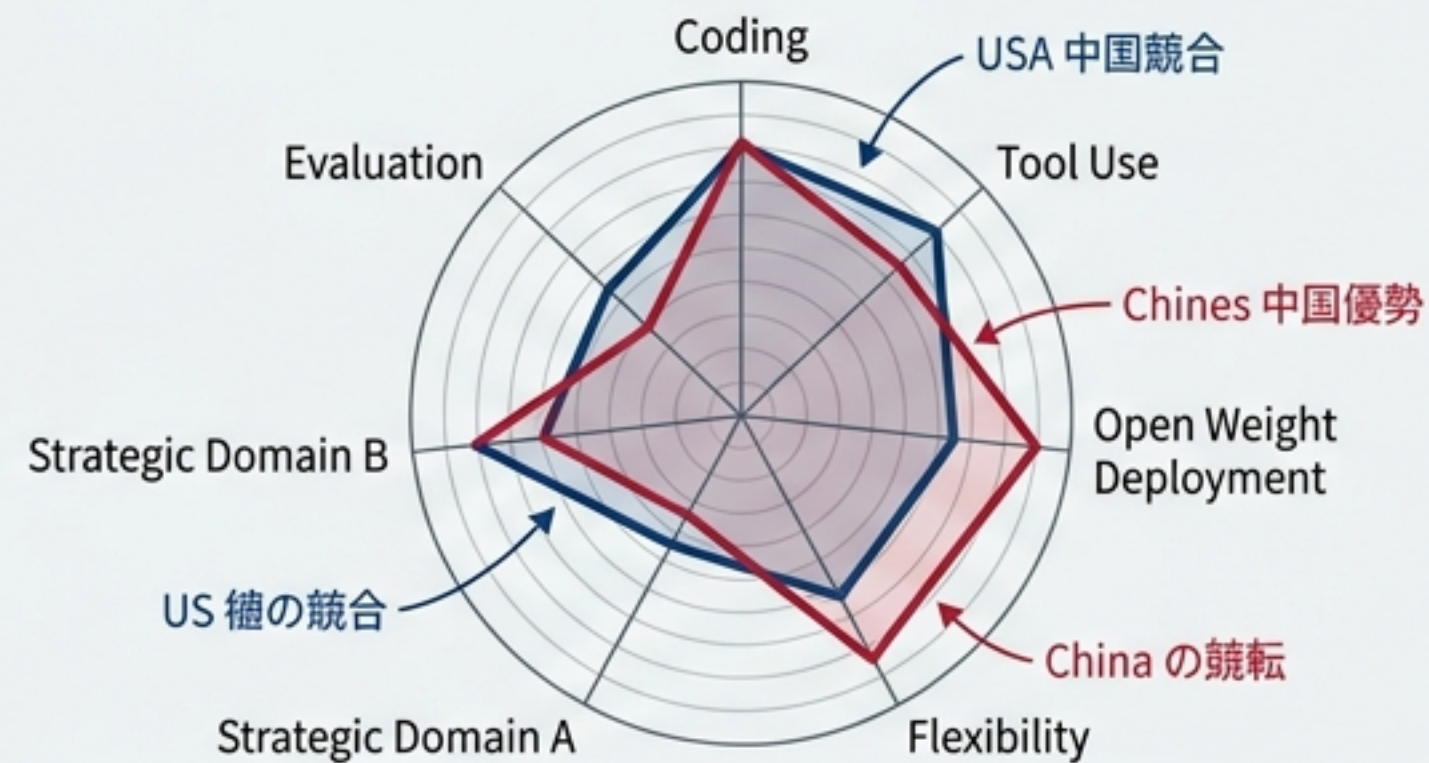
「中国勢は常に半年遅れる」という固定的な見方は、もはや成り立たない

従来の定説 (Myth)



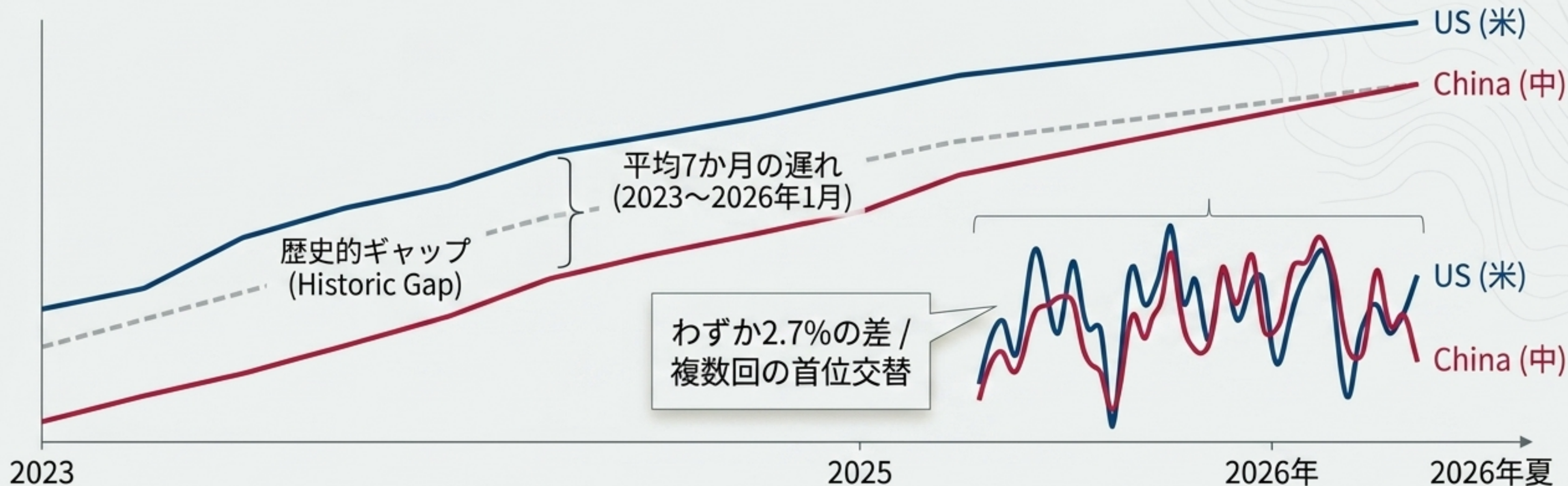
- 米国モデルが常に絶対的フロンティアを形成
- 中国モデルは、あらゆる能力において約6～7ヶ月遅れて追随
- 「国籍」がそのまま「性能階層」を意味する
- 「国籍」がそのまま「性能階層」を意味する

2026年夏の現実 (Reality)



- 同時期の競合: コーディング、ツール利用では同時期に競合・局所優位を獲得
- 提供形態の逆転: 中国勢はオープンウェイトを戦略的に展開し、配備の柔軟性で優勢
- 非一様な競争: 全能で米国を上回ったわけではなく、領域・評価・公開形態ごとに首位が交替する

外部評価の乖離は「長期的な歴史トレンド」と「直近の局所的拮抗」の違いに起因する



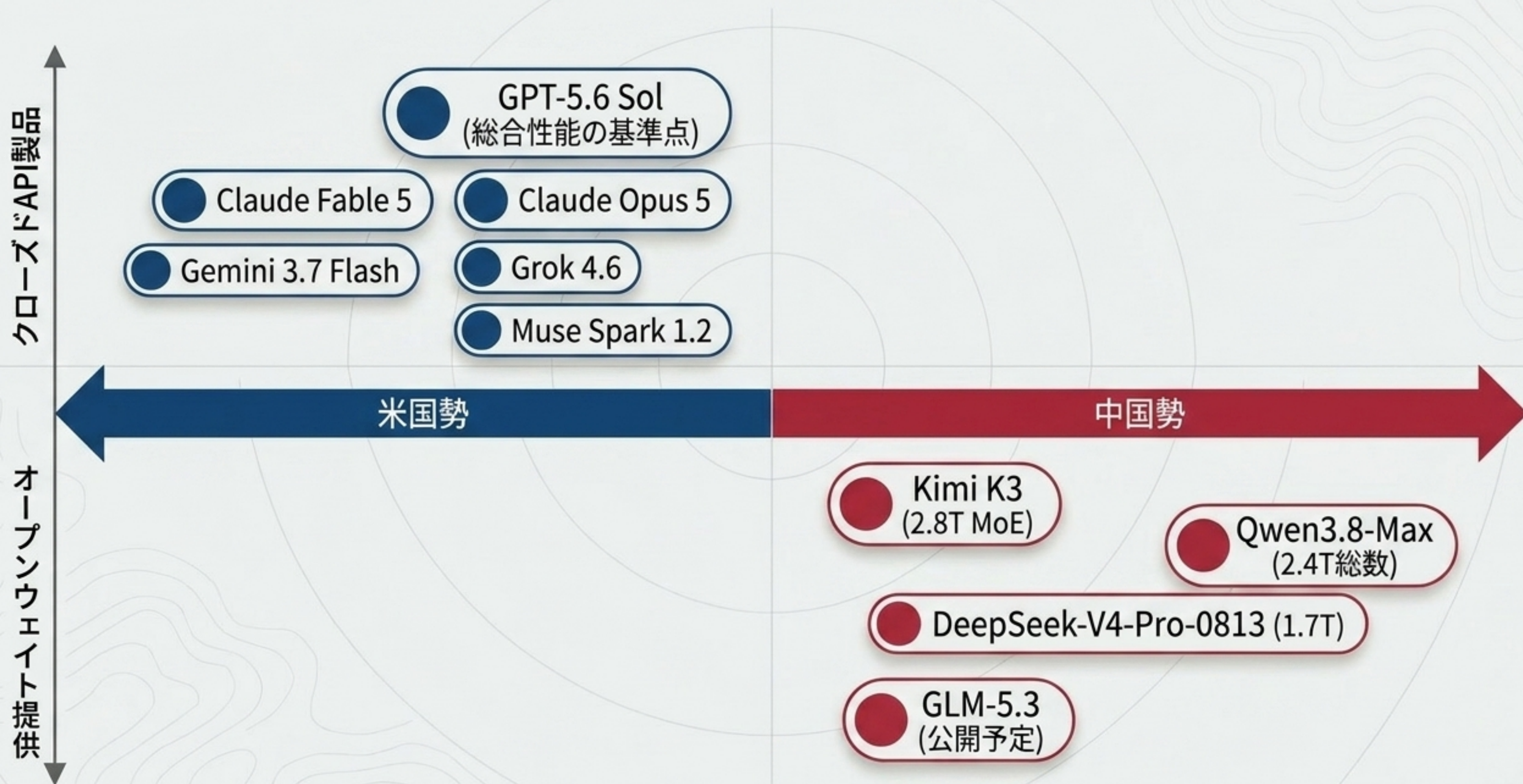
Epoch AI (2026年1月)

過去数年の能力フロンティアを一つの指数で追う長期的推計。当時の「半年程度の遅れ」という通念を定量的に裏付ける歴史的傾向。

Stanford HAI (2026年3月)

特定時点における総合的近接性。米中のモデル性能差は事実上閉じ、2025年初以降、首位が複数回交替しているリアルタイムの観測。

2026年夏のフロンティア10モデル：提供形態と能力による市場の二極化



米国勢（The Closed Titans）：最高峰の総合能力と成熟したエージェント製品環境

GPT-5.6 Sol



位置づけ:

GPT-5.6ファミリー最上位・API
主力閉鎖モデル

仕様:

約105万トークン文脈。既定4エ
ージェントを協調させる「ultra」
設計

指標:

AA Coding Agent Index 80 /
OSWorld 2.0 62.6%

Claude Fable 5



位置づけ:

一般提供Mythos級モデル。長期コ
ーディング・知識労働の最高水準

仕様:

100万トークン級。適応的思考、
長期エージェント機能。一部高リ
スク領域はOpus 4.8へフォールバ
ック

派生:

競合能力を半額で狙う主力「Opus
5」も展開

次世代ワークホース群

Gemini 3.7 Flash: 開発者・エ
ージェント利用主眼。DeepSWE
65.3%。低コスト提供。

Grok 4.6: 500K文脈。推論強度
とツール利用。

Muse Spark 1.2: ツール実行環
境と共同訓練。非同期・並列ツ
ール呼出し。

中国勢（The Open Challengers）：巨大パラメータの解放と局所的 特化

Kimi K3

仕様: 2.8Tパラメータ MoE（896中16専門家活性化）。ネイティブ視覚、100万トークン文脈

戦略: 総合性能でFable/Solに遅れを認める一方、長期コーディングと推論での優位を主張。重み公開済。

Qwen3.8-Max

仕様: 基礎公開版は2.4T総数・95B活性化。視覚入力、内蔵ツール追加。

戦略: Qwen-Max級として初のオープン公開。マネージド版と重み配布を並行。

DeepSeek-V4-Pro-0813

仕様: 1.7T、MITライセンス。DSpark投機的デコード、3段階の推論努力設定。

指標: Terminal Bench 87.9。重み配布済。

GLM-5.3

仕様: 長期コーディング / サイバー重視モデル。

指標: CyberGym 84.5（脆弱性発見に強み）。重みは安全評価後に公開予定。

収束する能力：エージェント型コーディングとツール利用における米中接近

Terminal Bench 2.1 (端末でのコーディング)



Insight: リポジトリ操作や端末利用を主用途とする場合、国籍だけで候補を除外する根拠は薄い。

Toolathlon Verified (ツール利用エージェント)



Insight: Kimiが米国上位モデルを局所的に上回る「首位交錯」状態。

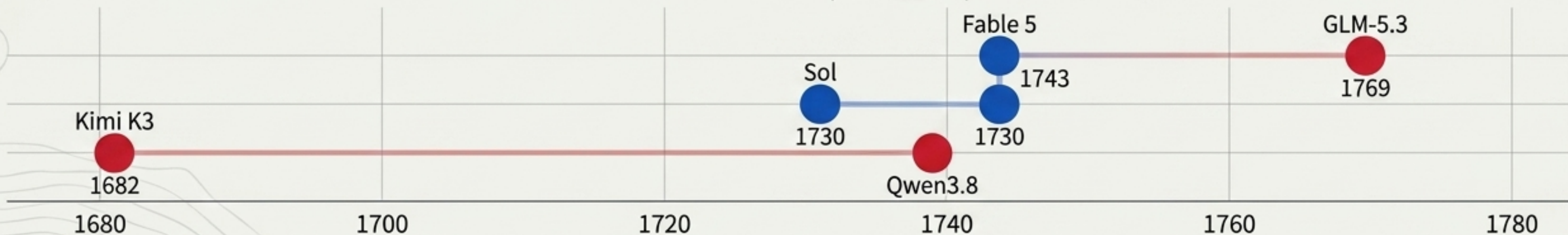
残されたフロンティア：長期の複合作業では米国最上位勢が「基準点」を維持

DeepSWE v1.1 (長期ソフトウェア工学)



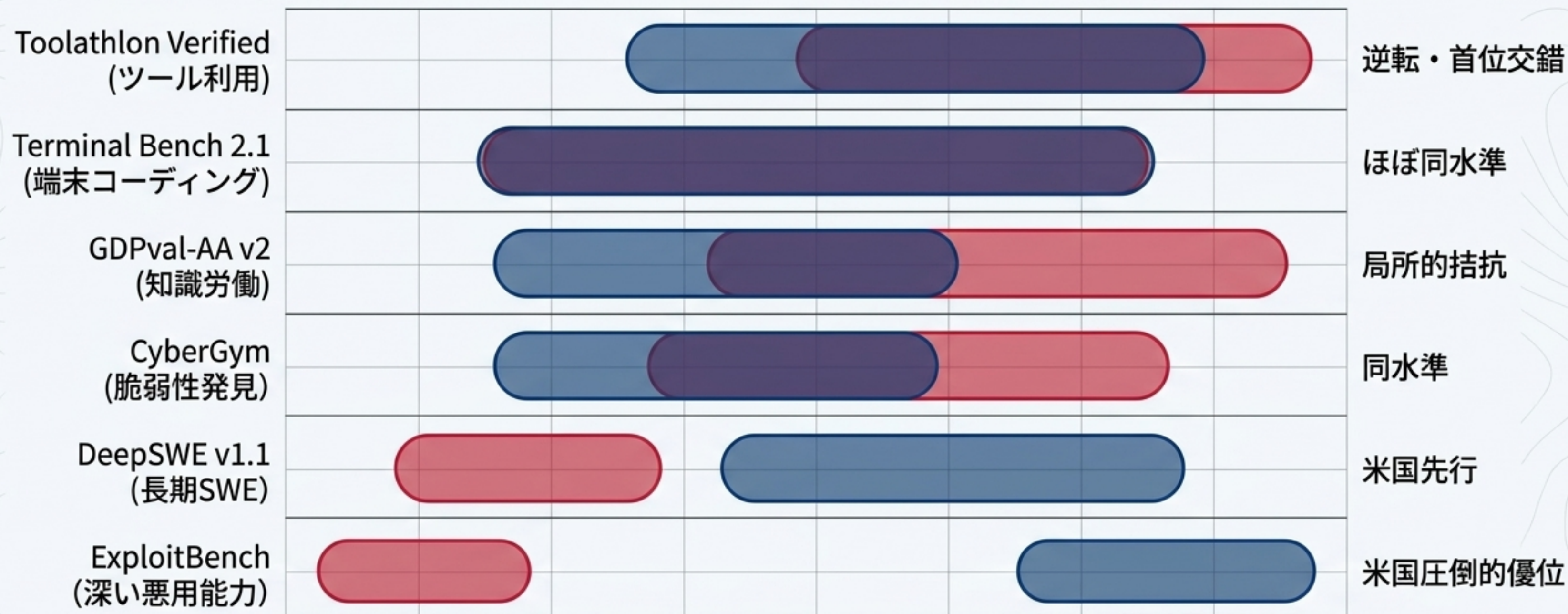
Insight: 長い実行軌跡、デバッグ、検証、環境フィードバックを必要とする評価では、Fable 5とSolが依然として圧倒的な優位を示す。Kimi K3の開発元自身も「総合性能ではこの両モデルにまだ届かない」と明記している。

GDPval-AA v2 (知識労働)



Insight: 一部の職務型評価（GLM等）で局所優位が見られるものの、全体としての安定性は米国勢に分がある。

6大指標ギャップ分析：性能差の「交錯」から「完全な乖離」まで

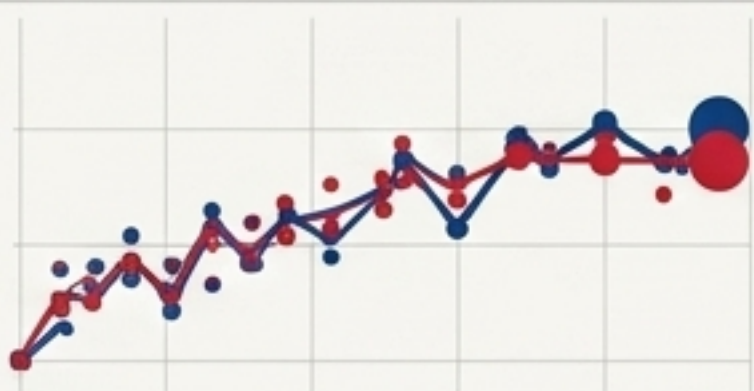


一本の時間軸による「遅れ」ではなく、評価対象の性質に応じた「非一様な競争」が視覚的に明らかである。

サイバー能力の解像度：「脆弱性発見」での拮抗と「深い悪用」での大差

フェーズ1: 発見と検証 (Discovery & Verification)

指標: CyberGym



スコア: GLM-5.3 (84.5)
vs Fable 5 (83.8)

状態: 【同等・局所優位】
GLMが一部で高い値を示す。

サイバー攻撃の複雑性チェーン (Cyber Complexity Chain)

フェーズ2: 深い悪用能力・自律実行 (Deep Exploitation & Execution)

指標: ExploitBench

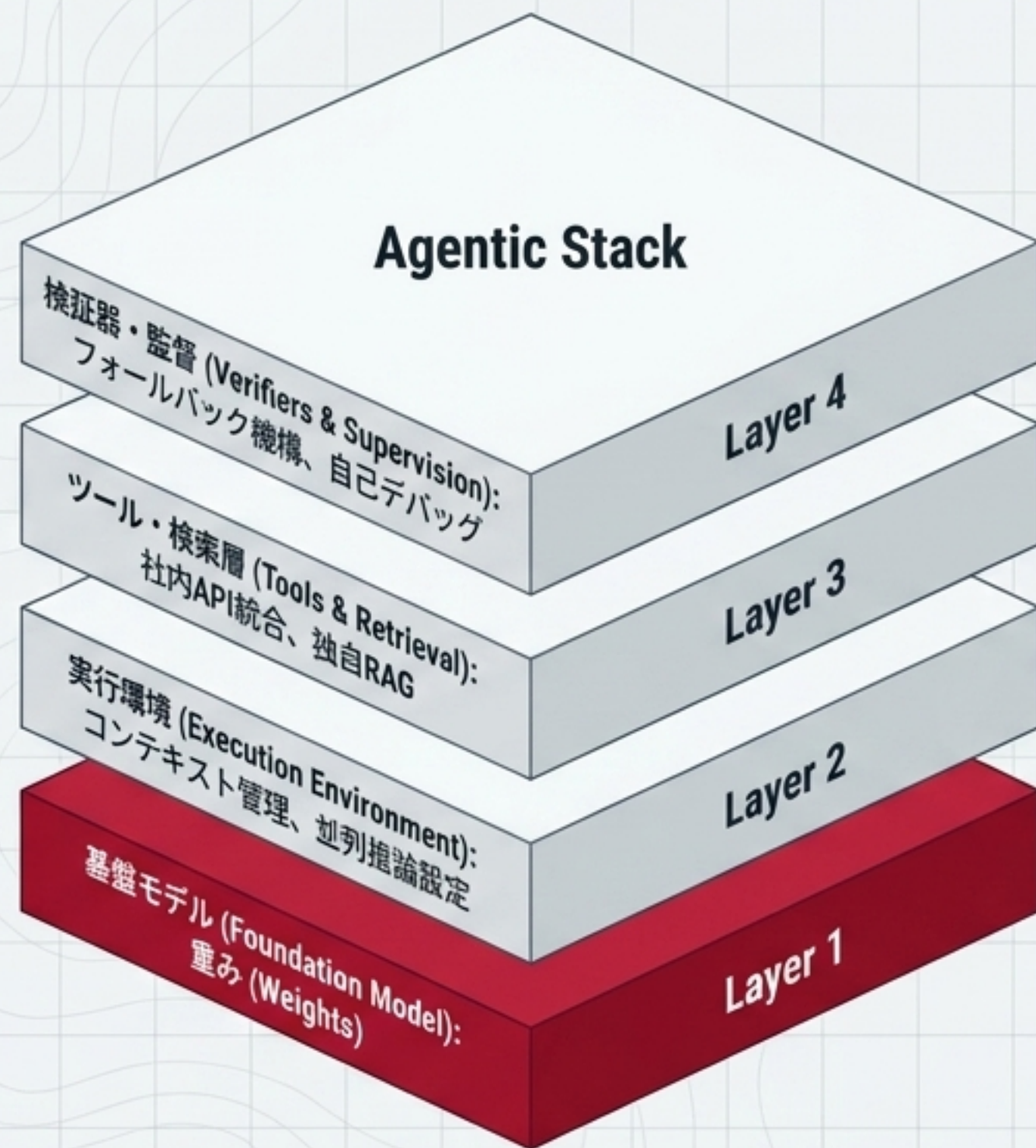


スコア: Fable 5 (78.0) / Sol (76.5)
vs GLM-5.3 (54.4) / Kimi K3 (32.2)

状態: 【大差での米国優位】
複雑な能力連鎖を伴う領域では、
依然として米国閉鎖モデルとの差
極めて大きい。

「サイバーで追いついた」という二分法は誤りである。
評価対象が発見か、検証か、悪用か、長期自律実行か
を示さない限り、能力評価は意味を持たない。

エージェント・ハーネスの支配：中国製オープンウェイトの真の戦略的価値



閉鎖型APIの限界:

米国モデルは最高峰の能力を持つが、ハーネス（上位層）はプロバイダの製品仕様に依存する。

オープンウェイトの優位性:

Kimi K3やQwen3.8の重み公開は、オンプレミス配備やデータ主権の確保だけでなく、「独自ハーネスへの深い最適化」と「地域・業界固有の運用」という生ベンチマークでは測れない実務価値を生む。

オープンウェイトの「安価で容易な利用」という錯覚

水面上の認識（見えるもの）

- ・モデルウェイトの無償ダウンロード
- ・サブスクリプション費用の回避
- ・ライセンスの自由（MIT等）

水面下の現実（隠れた障壁とコスト）

- ・超巨大インフラの要求: Qwenは2.4T、Kimiは2.8T、DeepSeekは1.7T級の巨大パラメータ。
- ・高度な運用ノウハウ: 量子化技術、推論ノードの並列化、遅延最適化。
- ・運用コスト: トークン消費・待ち時間を伴う推論努力・エージェント並列実行の計算コスト。

公開形態の優位は「小規模なPCで誰でも運用できる」ことではなく、「十分な計算基盤を持つ組織にとっての制御可能性と選択肢」にある。

総括：2026年 フロンティアAI 地政学的・領域別優位性マトリクス

	米国 (US)	中国 (China)
長期の複合作業 / 総合性能 (Long-term SWE)	圧倒的安定性 (Sol, Fable 5)	
複雑な悪用用・サイバー実行 (Deep Cyber Exploits)	米国製品の独壇場	
端末コーディング / ツール利用 (Terminal & Tools)	完全に拮抗・一部で局所優位	
データ主権・オンプレミス配備 (Data Sovereignty)		オープンウェイト展開
独自エージェントへの深い統合 (Custom Harness Integration)		フルスタックの制御権

モデル単体のベンチマークだけでなく、データセンター規模、製品統合、グローバル流通を含めると、最高性能を安定提供する「パッケージ」としては米国勢に依然としてアドバンテージがある。

エンタープライズ向け調達・導入フレームワーク

導入対象業務の最優先要件は何か？

【データ主権・自己ホスト・深い統合】

- 条件: データ所在の制約が厳しい / 独自基盤へ統合したい / 主用途が端末操作・コーディング
- 推奨候補: Kimi K3, Qwen3.8-Max, DeepSeek-V4, GLM-5.3 (将来)
- 留意点: 推論基盤と運用ノウハウへの自社投資が必須。

【最高品質・長期安定性・最小統合コスト】

- 条件: 複雑な長期作業が必要 / 成熟したエージェント環境を利用したい / インフラ投資を抑えたい
- 推奨候補: GPT-5.6 Sol, Claude Fable 5, Claude Opus 5
- 留意点: 入出力単価だけでなく、試行回数低下や監督工数削減によるトータルコストで評価する。

最終判断 (The Final Verdict)

01. 「半年遅れ」の終焉

「米国への半年遅れ」は歴史的傾向を表す近似としては有用だったが、現在の市場を説明する法則ではない。端末操作やツール利用において、中国勢は同世代の米国勢と完全に競合している。

02. 競争レイヤーの移行

競争は「純粋な推論性能」から「公開形態・ハーネスの柔軟性」へ移行。中国勢は1.7T超のオープンウェイト解放により、インフラを持つ組織に対してかつてない制御可能性を提供している。

03. 全能 vs 局所特化

米国勢は閉鎖型の最高峰モデルと統合製品群（データセンター、安全対策、グローバルサポート）で圧倒的な総合力と長期作業の安定性を保持。国籍ラベルによる調達排除は必須となる。