米中フロンティアモデル比較調査

対象: GPT-5.6 Sol、Claude Opus 5、Claude Fable 5、Gemini 3.7 Flash、Grok 4.6、Muse Spark 1.2、Kimi K3、Qwen3.8-Max、DeepSeek-V4-Pro-0813、GLM-5.3

調査基準日: 2026年8月15日（JST）

作成: Manus AI

エグゼクティブ・サマリー

本調査の結論は明確です。**「中国のフロンティアモデルは米国より常に半年遅れる」という固定的な見方は、2026年夏時点では成り立ちません。** 中国勢は、端末操作を伴うコーディング、ツール利用、長期エージェント実行、およびオープンウェイトの提供で、米国最上位モデルと同時期かつ一部の評価で競合・局所優位にあります。[8] [9] [10]

ただし、これは「中国勢が全能力で米国勢を一様に上回った」ことを意味しません。総合性能の基準点としては依然として **Claude Fable 5** と **GPT-5.6 Sol** が強く、Kimi K3の提供元自身も、総合性能ではこの両モデルにまだ届かないと説明しています。[1] [2] [7] 高難度の長期ソフトウェア工学、最高水準のエージェント推論、複雑なサイバー能力、および大規模な閉鎖型製品の運用成熟度では、米国勢の優位を示す証拠が残ります。

> **結論:** いま適切なのは「半年遅れ」という一本の時間差ではなく、**能力領域・ベンチマーク・公開形態ごとに首位が交替する非一様な競争**という理解です。中国勢はオープンウェイトと価格・配備の柔軟性で強く、米国勢は最高峰の総合能力と閉鎖型エージェント製品でなお優勢です。

論点	判断	根拠の要旨
固定的な半年遅れ	**支持しない**	Stanford HAIは2026年3月時点の米中差を2.7%とし、2025年初以降に首位交替が複数回あったと報告する。[11]
歴史的な遅れ	**限定的には支持**	Epoch AIは、2023年以降・2026年1月までのECIでは中国モデルの平均遅れを7か月（4〜14か月）と推定する。[12]
最新世代のコーディング／ツール利用	**中国勢は最前線に肉薄**	Terminal Bench、Toolathlon、GDPval-AA等では中国勢が米国上位勢と近接し、一部で上回る自社公表値がある。[8] [9] [10]
総合的最高性能	**米国勢が基準点**	GPT-5.6 SolとFable 5は、各社比較表・公開資料において広範な

		評価で最高水準に位置する。Kimi側もこの点を認める。[1][2][7]
配備・カスタマイズの自由度	**中国勢が優勢**	Kimi、Qwen、DeepSeekは重み公開済みで、GLMも公開予定。米国6モデルはAPI／自社製品中心の閉鎖提供である。[6][7][8][9][10]

調査の範囲と読み方

各モデルの開発元による公式発表、モデルカード、API文書、公式配布ページを一次資料として優先しました。性能値は、特記がなければ開発元の自己公表値です。とりわけエージェント評価は、エージェント・ハーネス、推論努力、ツール、最大出力、タイムアウト、試行回数、フォールバックの有無、評価版が異なります。従って、本稿ではスコアを**厳密な世界順位ではなく、能力の位置関係を示す証拠**として扱います。

例えばFable 5は、一部の高リスク領域でOpus 4.8へフォールバックする一般提供構成です。[2] また、GLM-5.3とQwen3.8の比較表は有用な横断情報を含む一方、競合値の一部に各社公表値や別ハーネスの最高記録を採用しています。[8][10] このため、数字の差が小さい場合は順位を断定しません。

1. 対象モデルの整理

米国フロンティアモデル

モデル	公表日	提供元・位置づけ	公表された主な仕様・提供形態	注目点
GPT-5.6 Sol	2026-07-09	OpenAIのGPT-5.6ファミリー最上位	テキスト・画像入力、約105万トークン文脈。`max`推論に加え、`ultra`では既定4エージェントを協調させる設計。APIの主力閉鎖モデル。	Agents' Last Exam 53.6、AA Coding Agent Index 80、BrowseComp 92.2%、OSWorld 2.0 62.6%をOpenAIが公表。[1]
Claude Fable 5	2026-06-09	Anthropicの一般提供Mythos級モデル	100万トークン級の長文脈、適応的思考、長期エージ	長期コーディング・知識労働・視覚で最高水準を主

			ェント、コード、視覚。サイバー／生物・化学等の一部要求はOpus 4.8にフォールバック。	張。Mythos 5は同一基盤を限定アクセス化した構成。[2]
Claude Opus 5	2026-07-24	Fable 5に近い能力を低価格で狙う主力モデル	API IDは `claude-opus-5`。入力$5/MTok、出力$25/MTok。Frontier-Bench、CursorBench、ARC-AGI 3、OSWorld等を重視。	Fable 5最高値の0.5%以内を約半額で達成したとする。Frontier-Benchは内部5試行平均であることを脚注で開示。[3]
Gemini 3.7 Flash	2026-08-13	Googleの「coding and agents」向けワークホース	100万トークン級のネイティブ・マルチモーダルモデル。API、Enterprise、Spark等に展開。年末まで入力$0.75／出力$3.75/MTok。	FrontierCode 43.6%、DeepSWE 65.3%、WebDev Arena Elo 1588を公表。低コストでの開発者・エージェント利用が主眼。[4]
Grok 4.6	2026-08-12	xAIの長期エージェント・対話／視覚制作向けモデル	API、Grok Build、Cursor等で提供。入力$2／出力$6/MTokから。500K文脈、推論強度とツール利用を提供。	AA Intelligence Index 61、GDPVal-AA v2 1753、CursorBench 69.9%を公表。競合値は各社資料・リーダーボード由来と明示。[5]
Muse Spark 1.2	2026-08-05	Metaのコーディング最適化モデル	100万トークン文脈、コンテキスト圧縮、非同期・並列ツール呼出し。Muse Code、Meta Model API、OpenRouterで提供。標準価格は入力$1.25／出力$4.25/MTok。	ツール実行環境と共同訓練し、長時間・複数エージェントのコーディングを設計中心に置く。数値入り比較の多くは図示である。[6]

中国フロンティアモデル

モデル	公表日	提供元・位置づけ	公表された主な仕様・公開形態	注目点
Kimi K3	2026-07-14	Moonshot AIのオープン・フロンティアモデル	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。重み公開とAPI／製品提供。	開発元は、総合性能ではFable 5とSolにまだ遅れると明記する一方、長期コーディング、知識労働、推論での競争力を主張。7
Qwen3.8-Max	2026-08-03	Alibaba Cloud／QwenのMax級公式版	基礎となる公開版は2.4T総数・95B活性化のMoE。Maxは視覚入力、非思考モード、標準100万トークン文脈、内蔵ツールを追加。	Qwen-Max級モデルとして初のオープン公開。Qwen Cloudでのマネージド版と重み配布を並行する。8
DeepSeek-V4-Pro-0813	2026-08-13	DeepSeek V4-ProのGA版	1.7T、MITライセンス、100万トークン文脈。DSpark投機的デコード、low/high/maxの推論努力、APIと重み配布。	HLE（ツールあり）60.0、Terminal Bench 87.9、DeepSWE 62.7を公式表で公表。DSBench系列は内部評価である。9
GLM-5.3	2026-08-14	Z.aiの長期コーディング／サイバー重視モデル	GLM-5.2と同一基盤をポストトレーニングで強化。low/high/maxの思考。Coding Plan・ZCodeで先行提供し、重みは安全評価後に公開予定。	CyberGym 84.5、DeepSWE 66.9、GDPval-AA v2 1769を公表。ExploitBenchではFable/Solとの差も併記している。10

2. 性能比較：どこで近づき、どこで差が残るか

2.1 指標別のスナップショット

次表は、GLM-5.3の公式比較表を中心に、同表が掲げる条件で読み取れる代表値を抜粋したものです。同表は、外部モデルの値に公開済みスコアを混在させるため、**同一実験によるランキングではありません**。しかし、「中国勢は全指標で大きく離されている」という見方がもはや正確でないこと、および差が残る領域を把握するには有益です。[10]

評価領域・代表指標	米国モデルの代表値	中国モデルの代表値	読み方
端末でのコーディング: Terminal Bench 2.1	GPT-5.6 Sol 88.8、Fable 5 88.0	Kimi K3 88.3、GLM-5.3 88.2、DeepSeek 87.9、Qwen3.8-Max 86.6	**ほぼ同水準**。ただしハーネスや設定は統一されない。[8][9][10]
長期ソフトウェア工学: DeepSWE v1.1	Sol 72.7、Fable 5 69.7	Kimi K3 67.5、GLM-5.3 66.9、DeepSeek 62.7、Qwen3.8-Max 56.6	中国勢は近接したが、同表では最上位米国勢が先行。[8][9][10]
ツール利用エージェント: Toolathlon Verified	Sol 74.9、Fable 5 74.7	Kimi K3 76.5、DeepSeek 74.1、GLM-5.3 73.0、Qwen3.8-Max 72.5	**首位が交錯**。Kimiはこの指標で高い値を示す。[9][10]
知識労働: GDPval-AA v2	Fable 5 1743、Sol 1730	GLM-5.3 1769、Qwen3.8-Max 1739、Kimi K3 1682	一部の職務型評価では中国勢の局所優位があり得る。[8][10]
脆弱性発見: CyberGym	Fable 5 83.8、Sol 83.6	GLM-5.3 84.5、DeepSeek 83.3、Kimi K3 80.0	発見・検証に限ればGLMが高い。安全上の含意と能力総体は別に評価する。[10]
より深い悪用能力: ExploitBench	Fable 5 78.0、Sol 76.5	GLM-5.3 54.4、Kimi K3 32.2、Qwen3.8-Max 28.8	複雑な能力連鎖では、GLM自身の比較表でも米国閉鎖モデルとの差が大きい。[10]

2.2 解釈

第一に、**エージェント型コーディングは収束が最も進んだ領域**です。Terminal Bench 2.1では、Kimi K3、GLM-5.3、DeepSeek-V4-Pro-0813がFable 5やSolとほぼ並ぶ自社公表値に達しています。[9][10] Qwen3.8-Maxも86.6と近接します。[8] このため、リポジトリ操作、端末利用、ツール呼出しを主用途とする導入では、国籍だけで候補を除外する根拠は薄くなっています。

第二に、**長期の複合作業では米国最上位勢がなお基準点**です。DeepSWEやFrontierSWEのように、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の優位を示す値が多く見られます。[1][8][10] ただしOpus 5やGemini 3.7 Flashは、最高性能のみ

を追うよりも、能力あたりのコストや速度を強く訴求しており、「最上位の一モデル」だけで市場を説明することも不十分です。[3][4]

第三に、**サイバー能力は単一尺度に還元できません**。GLM-5.3は脆弱性の発見・検証を測るCyberGymで高い値を示す一方、より深い悪用能力を問うExploitBenchではFable 5とSolに差が残ると自ら開示しています。[10] よって「サイバーで追いついた」あるいは「追いついていない」という二分法は、評価対象が発見、検証、悪用、長期自律実行のどれかを示さない限り意味を持ちません。

3. 競争構造を左右する公開形態

モデル能力だけでなく、**誰がどの形で使えるか**が競争優位を大きく左右します。指定された米国6モデルは、APIや自社製品を通じた閉鎖型提供が基本です。これに対し、中国勢ではKimi K3、Qwen3.8-2.4T-A95B、DeepSeek-V4-Pro-0813が重みを配布しており、GLM-5.3も公開予定を表明しています。[7][8][9][10]

この差は、モデルの生のベンチマーク性能だけでは測れない実務価値を生みます。オープンウェイトは、オンプレミス配備、データ主権への対応、自前の推論基盤、独自ハーネスへの深い最適化、地域・業界固有の運用に有利です。特にエージェントの品質は、基盤モデルだけでなく、ツール、検索、コンテキスト管理、検証器、実行環境に左右されるため、自由に構成を変えられる価値は大きいです。

一方、公開された重みが直ちに「安価で容易なローカル利用」を意味するわけではありません。Qwenは2.4T、Kimiは2.8T、DeepSeekは1.7T級であり、推論に必要なインフラ、量子化、並列化、運用ノウハウは依然として大きな障壁です。[7][8][9] したがって、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**にあります。

4. 「中国は半年遅れ」は払拭されたか

4.1 二つの外部評価は、見ている時点と定義が異なる

Epoch AIは2026年1月に、Epoch Capabilities Indexを用いて「2023年以降、中国モデルは米国フロンティアに平均7か月、最小4か月、最大14か月遅れた」と推計しました。[12] この分析は、過去数年の能力フロンティアを一つの指数で追うものであり、当時の「半年程度の遅れ」という通念には定量的な裏付けがあります。

他方、Stanford HAIの2026 AI Indexは、米中のモデル性能差は「事実上閉じた」とし、2025年初以降に米中モデルが首位を複数回交替し、2026年3月時点の米国首位との差は2.7%だと報告します。[11] これは、指標、候補モデル、測定時点がEpochの分析と異なるため、両者は必ずしも矛盾しません。前者は2026年1月までの長期的な先行／追随の推計、後者は2026年3月時点の総合的な近接を示すものです。

4.2 2026年夏の最新モデル群から得られる判断

ご指定のモデル群は、米国勢が2026年6月から8月、中国勢が7月から8月に集中して出揃っています。**リリース時期そのものには、もはや半年程度の機械的なズレは見られません。** また、コーディング、端末利用、ツール利用などでは、中国勢が同世代の米国勢と競合する数値を持ち、オープンウェイトの選択肢ではむしろ優位です。[7][8][9][10]

しかし、固定的な遅れが消えたことと、能力の全面的な同等化は別です。最新の一次資料でも、Kimi K3は総合性能でFable 5とSolを下回ると認め、GLM-5.3もより高度な悪用能力で閉鎖型米国モデルとの開きがあることを示します。[7][10] さらに、最高性能のモデルを安定して提供するためのデータセンター、製品統合、安全対策、グローバル流通、企業向けサポートまでを含めると、モデル単体のベンチマークを超える差は残っています。

> **最終判断:** 「半年遅れ」は歴史的傾向を表す近似としては有用でしたが、現在の市場を説明する法則ではありません。より正確には、**中国勢はオープンウェイトとエージェント実行で同時代の競争者となり、米国勢は閉鎖型の最高峰モデルと統合製品で強みを保つ**、という構図です。

5. 実務上の示唆

研究・調達・導入を検討する際は、国別のラベルではなく、第一に対象業務の評価を行うべきです。端末操作を伴うコーディング、自己ホスト、データ所在の制約、独自エージェント基盤への深い統合が重要であれば、Kimi K3、Qwen3.8、DeepSeek-V4-Pro-0813、将来のGLM-5.3重みは有力な検討対象です。[7][8][9][10] 一方、最少の統合コストでの最高品質、複雑な長期作業の安定性、成熟した閉鎖型エージェント環境を優先するなら、GPT-5.6 Sol、Claude Fable 5、Claude Opus 5が基準候補になります。[1][2][3]

価格比較も入力・出力単価だけで決めるべきではありません。高性能モデルは試行回数、再プロンプト、ツール呼出し、監督工数を減らせる場合がある一方、推論努力・エージェントの並列実行はトークン消費と待ち時間を増やします。したがって、実際のタスクセットで**成功率、総実行時間、総トークン、人的レビュー時間、失敗の影響**を一体で測ることが必要です。

参考文献

[1] OpenAI, GPT-5.6: Frontier intelligence that scales with your ambition
[2] Anthropic, Claude Fable 5 and Claude Mythos 5
[3] Anthropic, Introducing Claude Opus 5
[4] Google, Introducing Gemini 3.7 Flash
[5] xAI, Introducing Grok 4.6
[6] Meta, Meet Muse Spark 1.2 and Muse Code
[7] Moonshot AI, Kimi K3 Tech Blog: Open Frontier Intelligence
[8] Qwen, Qwen3.8-2.4T-A95B model card

[9] DeepSeek, DeepSeek-V4-Pro-0813 model card
[10] Z.ai, GLM-5.3: Frontier Coding with Emergent Cyber Capabilities
[11] Stanford HAI, The 2026 AI Index Report
[12] Epoch AI, Chinese AI models have lagged the US frontier by 7 months on average since 2023