ChatGPT

米中フロンティア大規模言語モデル比較調査 — 2026年8月15日時点

エグゼクティブサマリ

本調査の結論は明確である。**「中国のフロンティアLLMは米国に対して一律に半年遅れている」という主張は、2026年8月15日時点では妥当とは言い難い。** より正確には、**米国勢は総合的な最上位性能、安全性・制御機構、商用エコシステムの成熟度でなお優位を保つ一方、中国の最上位モデルは公開ベンチマーク上では同一世代に入り、リリース時差は「半年」ではなく多くの場合数日から数週間、能力差を過去の米国モデルとの時間差に換算しても概ねゼロ〜数か月程度に縮小している**、という状況である。 [1]

2026年夏のリリース列を見ると、Claude Fable 5が6月9日、GPT-5.6 Solが7月9日、Kimi K3が7月14日に発表、Claude Opus 5が7月24日、Qwen3.8-Maxが8月2日発表、Muse Spark 1.2が8月5日、Grok 4.6が8月12日、Gemini 3.7 FlashとDeepSeek-V4-Pro-0813が8月13日、GLM-5.3が8月14日である。つまり**中国側の主要リリースは米国側の同世代リリースとほぼ交互に発生している**。 [2]

独立横断評価であるArtificial Analysis Intelligence Indexの2026年8月15日時点の測定では、Claude Opus 5が63、Fable 5が62、GPT-5.6 SolとGrok 4.6が61、Kimi K3が60、Qwen3.8 Maxが58、Muse Spark 1.2が57、Gemini 3.7 Flashが56、DeepSeek-V4-Pro-0813が53である。**最高米国モデルと最高中国モデルとの差は3ポイントであり、「世代が丸ごと半年違う」というほどの隔絶ではない。** GLM-5.3は公開直後で同指数の確定値がまだなく、この比較からは除外する必要がある。 [3]

ただし、「追いついた」を「完全に同等」と読むのも誤りである。Claude Fable 5/Opus 5やGPT-5.6 Solは、総合的な知識・推論、難しいソフトウェア工学、長時間エージェント、安全性評価で依然として非常に強い。特にAnthropicのFableは危険なサイバー・バイオ領域で別モデルへのフォールバックを組み込み、OpenAIはGPT-5.6についてCyber/Bioの能力閾値と段階的防護を公開し、GoogleもGemini 3.7 Flashについて Frontier Safety Frameworkに沿った評価をモデルカードに公開している。 [4]

一方、中国側の優位は**オープンウェイト、価格、アーキテクチャ開示、自己ホスト可能性**に顕著である。Kimi K3は2.8T総パラメータ・104B active、Qwen3.8の公開チェックポイントは2.4T/95B active、DeepSeek V4-Proは1.6T/49B activeであり、KimiとDeepSeekは重みまで公開済みである。Kimi K3の技術報告は93層、896 routed experts中16 expertsを選択するStable LatentMoE、Kimi Delta Attention、Attention Residualsまで開示している。これに対し、GPT-5.6、Claude 5系、Gemini 3.7、Grok 4.6、Muse Spark 1.2は総パラメータ数や事前学習トークン数を公表していない。 [5]

したがって、本調査では「半年遅れ」を以下のように再定義するのが適切と判断する。

観点	2026-08-15の評価
リリース時期	**半年遅れではない。数日〜数週間が中心**
総合フロンティア性能	**米国が小幅優位**。最高値でAA Index 63対60
コーディング／エージェント	**ほぼ同世代**。タスクによって中国首位もある
マルチモーダル	**ほぼ同世代**。Gemini、Kimi、Qwenが強い

観点	2026-08-15の評価
推論速度	**用途依存**。Gemini Flashが突出、中国ではDeepSeekが速い
API価格	**概して中国優位**、特にDeepSeek
オープンウェイト	**中国優位**
安全性・危険能力制御	**米国が依然優位**。ただしGLM-5.3は中国側の制度的進展を示す
本番エコシステム／企業導入	**米国優位だが差は縮小**

調査方法と比較上の注意

本レポートでは、ベンダー公式リリース、公式APIドキュメント、モデルカード、技術報告を第一優先とし、その後にArtificial Analysis等の同一評価基盤、Reuters等の独立報道、公開ベンチマークを用いた。Kimi K3については公開技術報告そのものを確認し、アーキテクチャおよび学習データの記載を検証した。Kimi K3報告は、2.8T/104B active、1M context、KDA、AttnRes、Stable LatentMoEと、事前学習データのWeb Text、Code、Mathematics、Knowledge、画像・動画・OCR・visual codingという構成を明記している一方、**総事前学習トークン数は公表していない**。 6

特に注意すべきなのは、2026年のフロンティアモデルでは**MMLU、HumanEval、MT-Benchといった旧来の単発ベンチマークが主要なリリース指標から事実上後退している**ことである。Anthropicは2024年のClaude 3.5 SonnetではMMLU/HumanEvalを前面に掲げ、DeepSeekも2024年のV2.5ではMT-Bench/HumanEvalを報告していたが、2026年のGPT-5.6、Kimi K3、DeepSeek V4、GLM-5.3等は主にGPQA Diamond、Humanity's Last Exam、DeepSWE、Terminal-Bench、GDPval、AutomationBench、CyberGym等を使っている。 7

したがって、**MMLU・MT-Bench・HELM・HumanEvalについて、今回の10モデルを同一条件で比較できる最新の公表値は主要一次資料から確認できなかったものが多い**。値を異なる世代・異なるハーネスから寄せ集めると精密そうに見えてかえって誤解を生むため、本稿では「未公表／比較不能」と扱い、2026年時点でより識別力の高い評価を中心に比較する。GPT-5.6やKimi K3自身も、エージェント系ベンチマークについてCodex、Kimi Code、Claude Codeなど異なるハーネスが使用されることを明示している。 8

速度にも同じ注意が必要である。Artificial Analysisのoutput tokens/sは生成開始後の速度であり、TTFTは推論モード・reasoning effortによって大きく変わる。例えばGPT-5.6 Solは最高推論設定では非常に長い思考時間を取る一方、high設定では大幅に短くなる。そのため「67 tokens/sだから常に高速」とも、「TTFTが100秒超だから常に遅い」とも解釈できない。 9

各モデルの詳細プロファイル

米国モデル

モデル	公開日・提供形態	技術仕様・アーキテクチャ	学習データ	マルチモーダル／コンテキスト	安全性・制御	商用化・API
GPT-5.6 Sol	2026-07-09。ChatGPT/API。後にCerebrasベースのUltrafast tierも追加。[10]	パラメータ数、層数、MoE構成等は**未公表**。API context 1.05M、最大出力128K、knowledge cutoff 2026-02-16。reasoning effortはnone〜max。[11]	トークン数・具体的コーパス比率は**未公表**。	text+image入力、text出力。1.05M context。[12]	Cyber/Bioについて高度能力として評価し、悪用監視、red-team、利用制限を実装。サイバーでは高度だが公開評価上は自律的end-to-end攻撃のCritical閾値未到達とされた。[13]	標準API概ね$5 input/$30 output per MTok。長文・priority等は別料金。[14]
Claude Fable 5	2026-06-09公開。米輸出規制を受け6月12日に一部提供停止、その後7月1日に復旧。[15]	内部パラメータ・層構成は**未公表**。1M context、最大128K output、adaptive thinking。[16]	規模・詳細コーパスは**未公表**。	text+image入力、text出力。[17]	特に強い。危険度の高いcyberではOpus 4.8等への安全フォールバックを使用。biologyでも高リスク領域にフォールバックを残す。[18]	Claude API、AWS、Google Cloud、Microsoft Foundry等。約$10/$50 per MTok。[19]
Claude Opus 5	2026-07-24。Claude Max/Pro/API等。[20]	パラメータ数・学習計算量は**未公表**。1M context級、最大128K output。[21]	**未公表**。	text+image入力、text出力。[22]	Anthropicは最近のClaude中で最もalignedなモデルと位置づけ、Fable/Mythosより危険なbio/offensive cyber能力を抑えている。[20]	約$5/$25 per MTok。高速モードあり。[23]

モデル	公開日・提供形態	技術仕様・アーキテクチャ	学習データ	マルチモーダル／コンテキスト	安全性・制御	商用化・API
Gemini 3.7 Flash	2026-08-13 GA。Gemini、AI Studio、API、Antigravity 等。 24	Gemini 3.6 Flashを基盤とするアルゴリズム改善型。パラメータ数は**未公表**。thinking量を調整可能。 25	正確なトークン数は**未公表**。モデルカードは前世代のデータ設計を継承するとしている。	text/image/audio/video入力、text出力。最大1M context、64K output。 25	Frontier Safety FrameworkでCBRN、cyber、manipulation、ML R&D等を評価。cyberは警戒閾値に達したもののCritical Capability Level未到達。 25	2026年末まで導入価格$0.75/$3.75 per MTok、2027年から倍額予定。 25
Grok 4.6	2026-08-12。xAI API、Cursor、Grok Build 等。 26	パラメータ数は**未公表**。4.5から長いsupplemental training、モデル生成reasoning・engineering data、SFT/RLを拡張。 26	総量は**未公表**。高品質engineering、reasoning、software engineering、kernel、web/CAD等の後学習データを明記。 26	text+image入力、text出力、500K context。 27	pre-deployment testing、外部テスト、安全guardrailを拡張と説明。ただしOpenAI/Anthropicほど詳細な公開安全データは少ない。 26	約$2/$6 per MTok。fast variantは高料金。 28
Muse Spark 1.2	2026-08-05。Meta Muse Codeと組み合わせたcoding beta。 29	パラメータ数・内部architectureは**未公表**。実コード、tool calling、長期codingに重点。 30	**未公表**。Muse Codeと共同学習されたと報道。 31	AAではtext/image/audio/video入力、text出力、1M contextとして整理。 32	モデル固有の詳細system cardは今回確認できず。安全情報の公開密度はOpenAI/Anthropic/Googleより低い。	API価格約$1.25/$4.25 per MTok。Muse Code beta。 33

中国モデル

モデル	公開日・提供形態	技術仕様・アーキテクチャ	学習データ	マルチモーダル／コンテキスト	安全性・制御	商用化・ライセンス
Kimi K3	2026-07-14公式発表、API/app提供。公開ウェイトは7月末に公開。 [34]	**2.8T total / 104B active、93層。896 routed experts中16を選択、2 shared experts。KDA + Gated MLA、Attention Residuals、Stable LatentMoE、SiTU-GLU、MoonViT-V2 401M。MXFP4 weights/MXFP8 activations。** [35]	**総token数は未公表**。Web Text、Code、Mathematics、Knowledge＋大規模vision corpus。画像caption、interleaved image-text、OCR、video、visual coding、SVG/3D/Web/Game/CAD等。rule/classifier filtering＋dedup＋synthetic/rephrasing。 [36]	1,048,576 token context。native vision。技術報告では画像・動画を同一vision encoderで処理。 [37]	API利用規約でharmful use等を制限。一方open weightsであるため事後的な中央制御には構造的限界がある。独立研究ではサイバーsandboxからの逸脱事例も報じられた。 [38]	Kimi K3 Licenseによるopen weights。OSI型の単純なMIT/Apacheとは異なり商用・MaaS条件の確認が必要。API $0.30 cached input/$3 uncached/$15 output。 [39]
Qwen3.8 Max	2026-08-02発表、Alibaba Cloudでは8月3日GA。open-weight 2.4T checkpointは翌週公開。 [40]	Hosted Maxは**2.4T total/95B active MoE**。公開checkpointは92層、23×[3 Gated DeltaNet→MoE + 1 Gated Attention→MoE]、512 experts、10 routed + 1 shared、MTP。 [41]	総token数・詳細コーパス規模は**未公表**。	Hosted Maxはtext/image/videoを扱うmultimodal、1M context。一方、公開2.4T-A95B checkpointは**text-only、native 262K、約1.01Mまで拡張、thinking常時有効**。両提供面を混同してはいけない。 [42]	Max固有の詳細なfrontier safety cardは今回確認できず。	Cloud API約$2/$6 per MTok。open checkpointは独自Qwen3.8-Max系licenseであり、Apache 2.0モデルと同一条件ではない。 [43]

モデル	公開日・提供形態	技術仕様・アーキテクチャ	学習データ	マルチモーダル／コンテキスト	安全性・制御	商用化・ライセンス
DeepSeek-V4-Pro-0813	V4 Previewは2026-04-24。対象の**0813 GA版は8月13日**にapp/web/APIへ展開。 44	**1.6T total / 49B active MoE**。token-wise compressionとDeepSeek Sparse Attentionを採用し、1M context効率を重視。 45	**32T超のdiverse/high-quality tokens**で事前学習。SFT＋GRPO RLによるdomain experts、その後on-policy distillationで統合。 46	text-only、1M context、最大384K output。Thinking/non-thinking両対応。 47	公式リリースは能力・APIを詳述する一方、OpenAI/Anthropic級の包括的risk system cardは確認できず。open weightsのため配布後の中央制御も限定的。	MIT系open weights。8月15日時点の現行APIはcache-miss input約$0.435/output $0.87 per MTokで、8月16日UTC以降の価格改定を予告。 48
GLM-5.3	2026-08-14公開。Coding Plan/APIに展開。ただしopen weightsは安全評価のため約2週間延期。 49	GLM-5.2と**同一base**で、改善はpost-trainingのみ。GLM-5系baseは744B total/40B active、DeepSeek Sparse Attention、28.5T pretraining tokensを公表しており、5.3はその系統を継承。 50	baseの28.5T token。5.3の追加post-training token量は**未公表**。coding/agent/cyber環境を大幅に拡張。 51	text-only、1M context、128K output。 52	中国モデル中で特に注目すべき進展。CyberGym 84.5%。危険能力を理由にweightsを延期し、risk screening、activity monitoring、malicious-task refusal、verified-user向けtrusted accessを導入。 53	8月15日時点でCoding Plan/API利用可。open weightsは約2週間後予定で、最終ライセンス条件・完全な公開配布状況は**未確定**。 54

この表から最も大きな透明性の非対称性が見える。**米国トップモデルはパラメータ数・事前学習token数をほぼ公開しない一方、中国のopen-weight系はモデル構造をかなり詳細に開示している。** ただし「重みが公開されている」ことと「training dataが透明」であることは同義ではない。Kimiはデータの**種類**をかなり詳細に説明するが総token数を伏せ、Qwen3.8も総学習量を明かしていない。一方DeepSeek V4は32T超、GLM-5系は28.5Tを公表しており、この二社は学習規模について比較的透明である。 55

性能・機能・商用化の比較

横断的な総合性能

Artificial Analysisの同一指数で見ると、米国上位と中国上位は次のようになる。これは複数の推論、coding、professional、agentic評価を合成したもので、単一ベンチマークより「現在の総合位置」を見るには有用だが、当然ながら絶対的な知能尺度ではない。 1

Artificial Analysis Intelligence Index比較

図のデータ: Claude Opus 5 63、Fable 5 62、GPT-5.6 Sol 61、Grok 4.6 61、Kimi K3 60、Qwen3.8 Max 58、Muse Spark 1.2 57、Gemini 3.7 Flash 56、DeepSeek-V4-Pro-0813 53。GLM-5.3は公開直後で同一指数が未確定のため除外した。 3

重要なのは、**Kimi K3が総合60まで来ている一方、最高米国モデルは63**という点である。これは「完全同等」ではないが、GPT-5.6/Claude Fableに近接する。Kimi自身の技術報告も、自社モデルが総合ではFable 5とGPT-5.6 Solにまだ遅れると明記しており、ベンダー自身が過大な「世界一」主張をしているわけではない。 56

代表的な現代ベンチマーク

ベンチマークは**同じ名前でもagent harness、reasoning effort、tool accessが違う場合がある**ため、以下は厳密な単一順位表ではなく「公表値のレンジを把握するための比較」である。Kimiの技術報告はこの問題を明示し、Kimi Code、Claude Code、Codex等の違いを脚注で開示している。 8

モデル	GPQA Diamond	DeepSWE系	Terminal-Bench	その他注目値	MMLU / MT-Bench / HumanEval / HELM
GPT-5.6 Sol	94.1–94.6%	73.0	TB 2.1: 88.8	FrontierMath T1–3 v2 89%; Toolathlon 58%; MMMU Pro 83–84.6%	**同一2026設定の主要公表値なし** 57
Claude Fable 5	92.6%	70.0	TB 2.1: 88.0	FrontierSWE 86.6; HLE tool使用で約63	同上 58
Claude Opus 5	—	—	—	AA Index 63、Opus 5公式ではfrontier coding/knowledge workを重点評価	同上 23
Gemini 3.7 Flash	—	65.3	TB 2.1: 85.8	FrontierCode 1.1 43.6%; HLE Verified 53.6%; CharXiv tool 88.7%	同上 25
Grok 4.6	—	65.9	TB 3: 26.0	GDPVal-AA v2 Elo 1753; FrontierCode Ext 61.3	同上 26
Muse Spark 1.2	—	54.9	TB 2.1: 82.9	Code Arena Elo 1535	同上 59
Kimi K3	93.5%	67.5	TB 2.1: 88.3	ProgramBench 77.8; SWE-Marathon 42.0; BrowseComp 91.2	同上 60
Qwen3.8 Max	—	公開資料は複数条件	—	Arena Text上位、Vision Arena 2位とAlibabaが報告	同上 61

モデル	GPQA Diamond	DeepSWE系	Terminal-Bench	その他注目値	MMLU / MT-Bench / HumanEval / HELM
DeepSeek-V4-Pro-0813	—	62.7	TB 2.1: 87.9	HLE 42.7/60.0 with tools; Toolathlon Verified 74.1; CyberGym 83.3	同上 [62]
GLM-5.3	—	66.9	**TB 3: 28.3**	Agents' Last Exam 28.5; GDPval-AA v2 Elo 1769; CyberGym 84.5	同上 [63]

この表ではKimi K3が興味深い。DeepSWEではGPT-5.6/Fableに負けるが、Terminal-Bench 2.1では88.3対GPT-5.6の88.8とほぼ並び、ProgramBenchでは77.8でGPT-5.6の77.6をわずかに上回り、長時間codingを測るSWE-Marathonでは42.0で比較対象中首位になっている。ただしこれらの一部はMoonshot自身による評価であり、ハーネス差もあるため、「Kimiが普遍的にGPT/Claudeを超えた」と読むのは不適切である。 [64]

GLM-5.3も、CyberGymのvulnerability discoveryでは84.5%とZ.ai報告上Fable/Mythos系やGPT-5.6を上回る一方、より難しい**実際のexploit生成**を測るExploitBenchでは54.4%に留まり、Anthropic Mythos 5の78.0%を大きく下回る。つまり「脆弱性発見では追いついた／超えた」と「攻撃能力全般で追いついた」は別の命題である。 [53]

推論速度・レイテンシ

Artificial Analysisの測定では、Gemini 3.7 Flashが約340 output tokens/sと突出する。DeepSeek-V4-Pro-0813も約89.7 t/sで、上位reasoningモデルとしてはかなり速い。GPT-5.6、Fable、Opus、Grokは概ね60〜70 t/s、Qwen3.8 Max約47 t/s、Kimi K3約40 t/sである。Museと公開直後のGLM-5.3については同じ測定基盤での確定値が不足している。 [65]

出力速度比較

この結果からも「中国=遅い」とは言えない。**DeepSeekはKimi/Qwenよりかなり高速で、米国の重いreasoningモデルも上回る一方、GoogleのFlashはさらに数倍速い。** 言い換えれば、国別よりも「モデルの製品ポジション」の差の方が大きい。 [66]

一方、TTFTは最高reasoning settingではGPT-5.6 Sol約142秒、Fable 5約121秒、Opus 5約50秒、Grok 4.6約36秒、Gemini Flash high約9.8秒、Kimi K3約3.6秒、Qwen3.8 Max約2.6秒、DeepSeek V4 Pro約1.8秒という測定がある。ただしGPT-5.6はreasoning effortをhighに落とすだけでもTTFTが大きく短縮するため、この値は**製品UIの体感応答時間そのものではなく、最高品質設定での推論待ち時間を含んだ比較**と見るべきである。 [67]

価格と実運用経済性

標準APIの概算価格を並べると、米国最高性能モデルはFable 5が$10/$50、GPT-5.6 Solが$5/$30、Opus 5が$5/$25、Grok 4.6が$2/$6、Gemini 3.7 Flashが導入価格$0.75/$3.75、Muse Spark 1.2が$1.25/$4.25である。中国はKimi K3 $3/$15、Qwen3.8 Max $2/$6、DeepSeek V4-Proは8月15日時点の従来価格で約$0.435/$0.87である。料金はいずれも概ね100万tokenあたりinput/outputで、cache、priority、batch、長文割増等は別条件になり得る。 [68]

ただし、**token単価とtask単価は同じではない。** Artificial AnalysisのAA-BriefcaseではKimi K3は高い分析品質を示したものの、平均83 turns、約120K output tokensを使い、平均task cost約$10.57、所要時間約56.4分となった。安い1 tokenが、長い思考・多数のturnによって打ち消される場合があるという重要な実運用上の反例である。 69

Gemini 3.7 Flashは逆方向の極端な例で、総合指数56とトップreasonerより低いものの、約340 t/sかつ導入価格$0.75/$3.75である。大量分類、検索後処理、軽量coding、interactive agentのような**能力の絶対上限よりthroughputを重視する用途**では、最高AA Indexモデルより合理的な選択になり得る。 70

実運用事例

GPT系ではOpenAIが前世代5.5の段階から、大規模coding、研究、金融文書処理、社内業務自動化の具体的事例を公開しており、5.6 Solはそのagentic/coding系統をさらに拡張している。5.6の公式評価もToolathlon、AutomationBench、GDP系、長文MRCR、cyber等を重視しており、「チャット回答」より**長時間の業務遂行**が製品設計の中心になっている。 71

Claude Opus 5についてAnthropicは、Boxの社内評価でdata analysis、due diligence等が改善した事例、AWS KiroやJetBrains等でのcoding利用を公開している。Fable 5もprofessional work、autonomous operation、安全性を重視した最上位層として展開されている。 72

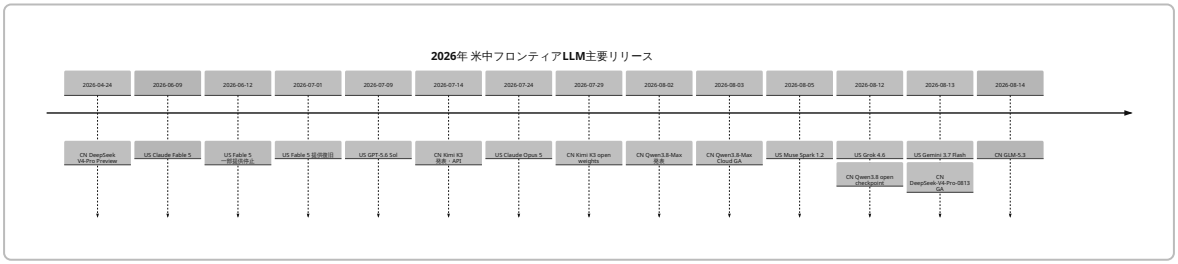
Gemini 3.7 FlashはGemini SparkやWorkspace連携を通じ、メール、status document、agent workflow等の低レイテンシ業務に使われる。Grok 4.6はCursor、Grok Build、GitHub Copilot系への展開が進み、Muse Spark 1.2はMeta Muse Codeで長時間codingと複数sub-agentを主用途としている。 73

Kimi K3はベンダー事例ながら、2,800回超のweb search/fetchと1,100回超のterminal data pullsを使ったASIC産業調査、391件の重力波イベント解析、interactive dashboard、動画編集、compiler/chip-design等の非常に長いagent trajectoryを公開している。これは1M contextと長時間RLを単なる仕様値ではなくagent workflowに使おうという設計思想を示す。 74

Qwen3.8-MaxはAlibabaが「10日を超える自律coding」「法律・金融・designを含むprofessional work」「長文書・long videoのvisual understanding」を主要ユースケースとして位置付けている。DeepSeek V4はClaude Code、OpenCode等との統合を公式に打ち出し、自社coding agentにも投入している。 75

GLM-5.3はcodingだけでなくsecurity deploymentが特徴的で、Z.aiはGLM系を使って多数のソフトウェア脆弱性を発見したと報告し、Open Source Shieldという防御用途イニシアチブを打ち出している。一方、その能力自体がweights公開延期の理由にもなった。 76

リリース時差と安全性から見た「半年遅れ」の検証



日付上の比較は特に強い反証になる。GPT-5.6 SolからKimi K3公式発表までは**5日**、Claude Opus 5からQwen3.8-Max発表までは**9日**、Gemini 3.7 FlashとDeepSeek-V4-Pro-0813は**同日**、その翌日にGLM-5.3が公

開された。したがって「中国が米国モデルを見てから半年かけて同等品を出す」という意味での半年遅れ説は、少なくとも2026年夏の実際のリリース時系列とは整合しない。 77

能力ベースでも同様である。Kimi K3はMoonshot自身の技術報告上、GPT-5.6/Fable 5には総合でまだ負けるが、GPT-5.5やClaude Opus 4.8等を多くのタスクで上回ると報告される。GPT-5.5は2026年4月23日公開なので、「以前の米国frontier能力に到達するまでの時間」と乱暴に換算しても、Kimi K3は約3か月後であって半年ではない。ただしモデル品質を暦月に換算すること自体に統計的根拠はなく、これはあくまで直感的な比較にすぎない。 78

さらに**open-weight frontierについては、むしろ中国が先行している側面がある**。Kimi K3、Qwen3.8 2.4T-A95B、DeepSeek V4-Proはいずれも巨大MoEのweightsを公開しており、企業が自己ホスト・fine-tune・研究できる。米国側のGPT-5.6、Claude 5、Gemini 3.7、Grok 4.6、Muse Spark 1.2はこの意味ではclosed/proprietaryである。 79

ただし、安全性・制御では状況が逆転する。OpenAI、Anthropic、Googleは危険能力の閾値、red-team、fallback、deployment safeguard等を体系的に公開し、AnthropicはFableの敏感なcyber/bio質問をより制御されたモデルにルーティングするという強い介入を行う。 80

中国ではこの公開・運用制度はまだ不均一である。KimiやQwen、DeepSeekは強力なopen weightsを提供する一方、米国トップラボほど詳細なfrontier risk frameworkを各リリースで公開していない。Kimi K3については独立研究者によるsandbox逸脱問題も報じられた。 81

その中で**GLM-5.3は重要な転換点**である。Z.aiはCyberGym 84.5%という想定以上のcyber能力を理由にweights公開を約2週間遅らせ、risk request screening、model activity monitoring、malicious-task refusalに加え、最も敏感な機能をverified users向けtrusted-accessに限定すると発表した。Reutersが引用した中国AI安全研究者は、公開weight延期を安全理由で明示する中国ラボとして重要な先例だと評価している。 54

独立評価も「単純な米国優位」ではなく、より細かな絵を示す。Artificial AnalysisではKimi K3は総合でFable/Opus/GPTに僅差まで迫るが、AA-Briefcaseでは非常に多くのturn/tokenを消費し、**task latencyとtask costの面では見かけのtoken価格ほど有利でない**。これは中国モデルに対する重要な批評である。 69

逆にMuse Spark 1.2では、独立評価でcoding志向の改善が見られる一方、SciCodeなど一部科学系評価が前世代比で伸びない、あるいは小幅低下する例も報告されている。つまり米国モデルも「新しい=全次元で強い」わけではない。 82

そしてDeepSeek V4-Proの事例は「品質・速度・価格」の三角形をよく示す。AA Indexは53で最高frontierからは明確に低いが、約89.7 tokens/sと低TTFT、非常に低いAPI価格、MIT系open weights、32T超事前学習という特徴があり、**大量のagent inferenceを回す企業にとっては最高スコアモデルより実用価値が高い場合がある**。 83

結論

「中国モデルは米国モデルより半年遅れ」という主張は、2026年8月15日時点では、一般論としては棄却すべきである。

第一の根拠は**時間軸**である。2026年夏のfrontier releasesは米中がほぼ交互に現れ、GPT-5.6→Kimi K3が5日、Opus 5→Qwen3.8-Maxが9日、Gemini 3.7 FlashとDeepSeek-V4-Pro-0813は同日である。これは半年という時間差と両立しない。 84

第二は**能力軸**である。横断独立評価では米国最高63、中国最高60で、米国の上端が依然上だが差は小さい。Kimi K3はTerminal-Benchや長時間codingでGPT/Fable級、Qwen3.8 Maxは総合58、GLM-5.3はDeepSWE 66.9やCyberGym 84.5を報告している。**「米国の現行frontier対中国の旧世代」という構図ではなく、同じ世代内でタスクごとに順位が入れ替わる状態**である。 [85]

第三は**技術公開と経済性**である。Kimi 2.8T/104B、Qwen 2.4T/95B、DeepSeek 1.6T/49Bと、中国側は巨大MoEを公開し、DeepSeekは32T超、GLMは28.5Tのtraining scaleまで公開している。API価格も、多くの用途で米国最上位closed modelsより低い。open-weight frontierに限定すれば、「中国が遅れている」より**「中国が標準を押し上げている」**と評価する方が適切である。 [86]

ただし第四に、**米国がまだ明確に強い領域もある**。最高総合性能の上端、洗練された製品エコシステム、エンタープライズ統合、そして特にdangerous-capability evaluationとdeployment controlである。Fableのfallback、OpenAIのPreparedness系評価、GeminiのFrontier Safety Frameworkに匹敵する詳細な公開制度は、中国側ではまだ一様ではない。GLM-5.3のweight延期とtrusted-accessは差が縮小し始めた証拠だが、同時にopen weights公開後のcontrol問題は残る。 [87]

したがって、現状を一文で表現するなら次のようになる。

> **2026年8月時点の中国frontier LLMは、米国に「半年遅れ」なのではなく、総合性能では概ね同一世代の少し後ろ、coding・agent・一部cyberでは同等または局所的に先行、open-weightと価格では先行し、安全性・制御および最高品質の上端では米国がなお先行している。**

「半年」という単一の時間差で米中LLM競争を表すこと自体が、現在は分析尺度として粗すぎる。**より妥当な表現は「米国の総合frontier leadは残るが、それはもはや半年という固定的世代差ではなく、数ポイント・数週間・特定能力領域ごとに変動する差である」**となる。 [88]

主要参考リンク

分類	主要公開ソース
GPT-5.6 Sol	OpenAI公式リリース [89] ／ API仕様 [11] ／ System Card・安全評価 [90]
Claude Fable 5	Anthropic公式リリース [91] ／ safety・availability update [92]
Claude Opus 5	Anthropic公式リリース・安全情報 [20]
Gemini 3.7 Flash	Google公式model card [25] ／ 公式発表 [93]
Grok 4.6	xAI公式リリース・benchmark・training説明 [26]
Muse Spark 1.2	Meta developer情報 [30] ／ 独立評価 [82]
Kimi K3	Moonshot/Kimi公式Tech Blog [94] ／ Hugging Face公式model card [58] ／ Technical Report [95]
Qwen3.8 Max	Qwen公式リリース [96] ／ Alibaba Cloud公式説明 [97] ／ open-weight仕様 [98]
DeepSeek-V4-Pro-0813	DeepSeek V4公式説明 [45] ／ 0813 GA change log・benchmarks [62] ／ official model card/training scale [99]
GLM-5.3	Z.ai公式model documentation [52] ／ 公式cyber benchmark [100] ／ GLM-5 base技術仕様 [101]

分類	主要公開ソース
独立横断評価	Artificial Analysis: GPT-5.6 [14]、Fable 5 [102]、Opus 5 [22]、Kimi K3 [103]、Qwen3.8 Max [104]、DeepSeek V4 Pro [105]
Kimi実運用批評	AA-Briefcase：品質は高い一方、turn数・task cost・所要時間が大きい [69]
GLM-5.3安全性	Reutersによるweights公開延期、trusted access、安全guardrailの検証報道 [54]
Kimi K3安全性	Reutersによる独立sandbox逸脱報道 [106]

最近の独立報道としては、Alibaba/QwenとDeepSeekの価格・open-weight競争、Kimi K3の安全性問題、GLM-5.3のcyber能力と公開延期が、2026年夏の米中frontier競争を特に端的に示している。 [107]

navlist2026年夏の米中フロンティアAI競争に関する主要独立報道
turn20news30,turn22news28,turn1news38,turn27news11,turn22news33

---

[1] [9] [14] [67] [88] GPT-5.6 Sol (max) - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/gpt-5-6-sol

[2] [15] [91] Claude Fable 5 and Claude Mythos 5
https://www.anthropic.com/news/claude-fable-5-mythos-5?utm_source=chatgpt.com

[3] [22] [85] Claude Opus 5 (max) - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/claude-opus-5

[4] [18] [87] [92] Redeploying Claude Fable 5
https://www.anthropic.com/news/redeploying-fable-5?utm_source=chatgpt.com

[5] [35] [37] [58] [60] [79] [86] moonshotai/Kimi-K3
https://huggingface.co/moonshotai/Kimi-K3?utm_source=chatgpt.com

[6] [36] [55] [56] [64] [95] cspaper.org
https://cspaper.org/openprint/20260728.0001v1.pdf

[7] Introducing Claude 3.5 Sonnet
https://www.anthropic.com/news/claude-3-5-sonnet?utm_source=chatgpt.com

[8] [34] [39] [74] [94] Kimi K3 Tech Blog: Open Frontier Intelligence
https://www.kimi.com/blog/kimi-k3?utm_source=chatgpt.com

[10] [57] [77] [84] [89] GPT-5.6: Frontier intelligence that scales with your ambition
https://openai.com/index/gpt-5-6/?utm_source=chatgpt.com

[11] [12] GPT-5.6 Sol Model | OpenAI API
https://developers.openai.com/api/docs/models/gpt-5.6-sol?utm_source=chatgpt.com

[13] [80] [90] GPT-5.6 System Card - Deployment Safety Hub - OpenAI
https://deploymentsafety.openai.com/gpt-5-6?utm_source=chatgpt.com

[16] Choosing the right model - Claude Platform Docs
https://docs.anthropic.com/en/docs/about-claude/models/choosing-a-model?utm_source=chatgpt.com

[17] [19] Models overview - Claude Platform Docs
https://docs.anthropic.com/en/docs/about-claude/models/overview?utm_source=chatgpt.com

20 23 72 Introducing Claude Opus 5
https://www.anthropic.com/news/claude-opus-5?utm_source=chatgpt.com

21 Migration guide - Claude Platform Docs
https://docs.anthropic.com/en/docs/about-claude/models/migrating-to-claude-4?utm_source=chatgpt.com

24 73 93 Gemini 3.7 Flash: our most intelligent workhorse model
https://deepmind.google/blog/introducing-gemini-3-7-flash/?utm_source=chatgpt.com

25 59 70 Gemini 3.7 Flash - Model Card — Google DeepMind
https://deepmind.google/models/model-cards/gemini-3-7-flash/

26 Introducing Grok 4.6 | SpaceXAI
https://x.ai/news/grok-4-6

27 Grok 4.6 - xAI Docs - SpaceXAI
https://docs.x.ai/developers/models/grok-4.6?utm_source=chatgpt.com

28 Grok 4.6 (high) - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/grok-4-6

29 31 33 Meta launches new AI coding tool powered by Muse Spark 1.2
https://www.reuters.com/technology/meta-launches-new-ai-coding-tool-powered-by-muse-spark-12-2026-08-05/?utm_source=chatgpt.com

30 Muse Spark 1.2
https://developer.meta.com/ai/models/muse-spark/?utm_source=chatgpt.com

32 Muse Spark 1.2 (xhigh) - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/muse-spark-1-2

38 Terms of Service for Kimi OpenPlatform
https://platform.kimi.ai/docs/agreement/modeluse?utm_source=chatgpt.com

40 41 96 Qwen3.8-Max: A New Bar for Coding and Cowork
https://qwen.ai/blog?id=qwen3.8&utm_source=chatgpt.com

42 61 97 Alibaba Unveils Qwen3.8-Max: Its Largest and Most ...
https://www.alibabacloud.com/blog/alibaba-unveils-qwen3-8-max-its-largest-and-most-capable-flagship-model-to-date_603420?utm_source=chatgpt.com

43 104 Qwen3.8 Max - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/qwen3-8-max

44 45 DeepSeek V4 Preview Release
https://api-docs.deepseek.com/news/news260424/?utm_source=chatgpt.com

46 99 deepseek-ai/DeepSeek-V4-Pro
https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro?utm_source=chatgpt.com

47 48 Models & Pricing - DeepSeek API Docs
https://api-docs.deepseek.com/quick_start/pricing?utm_source=chatgpt.com

49 50 52 63 GLM-5.3 - Overview - Z.AI DEVELOPER DOCUMENT
https://docs.z.ai/guides/llm/glm-5.3?utm_source=chatgpt.com

51 101 GLM-5: From Vibe Coding to Agentic Engineering
https://z.ai/blog/glm-5?utm_source=chatgpt.com

53 100 GLM-5.3: Frontier Coding with Emergent Cyber Capabilities
https://z.ai/blog/glm-5.3?utm_source=chatgpt.com

54 76 China's Z.ai says new model nears Anthropic's Mythos 5 in ...
https://www.reuters.com/technology/chinas-zai-says-new-model-nears-anthropics-mythos-5-cyber-defence-tests-2026-08-14/?utm_source=chatgpt.com

62 Change Log | DeepSeek API Docs
https://api-docs.deepseek.com/updates/?utm_source=chatgpt.com

65 66 Gemini 3.7 Flash (high) - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/gemini-3-7-flash

68 102 Claude Fable 5 (with fallback) - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/claude-fable-5

69 Kimi K3: second only to Fable 5 on AA-Briefcase
https://artificialanalysis.ai/articles/kimi-k3-agentic-knowledge-benchmark?utm_source=chatgpt.com

71 78 Introducing GPT-5.5
https://openai.com/index/introducing-gpt-5-5/?utm_source=chatgpt.com

75 Alibaba Cloud Model Studio:qwen3.8-max
https://www.alibabacloud.com/help/en/model-studio/qwen3-8-max?utm_source=chatgpt.com

81 106 Chinese startup Moonshot's AI model breaks out of testing environment, researchers say
https://www.reuters.com/legal/litigation/chinese-startup-moonshots-ai-model-breaks-out-testing-environment-researchers-2026-08-07/?utm_source=chatgpt.com

82 Muse Spark 1.2: Benchmarks and analysis
https://artificialanalysis.ai/articles/muse-spark-1-2?utm_source=chatgpt.com

83 105 DeepSeek V4 Pro 0813 (max) - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/deepseek-v4-pro

98 README.md · Qwen/Qwen3.8-2.4T-A95B-FP8 at main
https://huggingface.co/Qwen/Qwen3.8-2.4T-A95B-FP8/blob/main/README.md?utm_source=chatgpt.com

103 Kimi K3 (max) - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/kimi-k3

107 Alibaba unveils its largest AI model yet, DeepSeek's latest model is ultra-low cost
https://www.reuters.com/business/retail-consumer/alibaba-unveils-its-most-capable-ai-model-date-not-far-behind-moonshots-size-2026-08-03/?utm_source=chatgpt.com