ChatGPT

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

エグゼクティブサマリ

調査基準日は2026年8月16日（日本時間）である。今回の対象10モデルを、公式発表、モデルカード、API仕様、公開ウェイト、独立評価のArtificial Analysis、主要報道などで突き合わせた結果、結論はかなり明確になった。

第一に、「中国のフロンティアAIは米国より一律に半年遅れ」という表現は、2026年8月時点ではもはや妥当ではない。 独立評価Artificial Analysisの現行Intelligence Indexでは、米国側上位がClaude Opus 5の63、Claude Fable 5の62、GPT-5.6 SolとGrok 4.6の61であるのに対し、中国のKimi K3は60、Qwen3.8 Maxは58に達している。特にKimi K3は**オープンウェイトで60**であり、最上位の米国クローズドモデルとの差は2〜3ポイントに縮まった。これは「半年」という時間差で説明できる状態ではない。むしろ、**モデル能力では“ほぼ同一世代”、特定分野では相互に抜きつ抜かれつ**と見るべきである。 [1]

ただし、**「中国が全面的に追いついた」という結論も早い。** 米国には、NVIDIA Rubinを中心とする最先端アクセラレータ、CUDA、巨大クラウド、超高速推論インフラ、Cerebras等を組み合わせたフルスタック優位が残る。Rubinはすでに量産段階に入り、2026年後半からパートナー製品が供給される。一方、中国は米国の半導体輸出規制の制約下にあり、Huawei Ascend系と巨大SuperNode、MoE、低精度化、スパースAttentionなどを組み合わせて**「計算資源の不利をアルゴリズムとシステムで相殺する」**方向に進んでいる。 [2]

第二に、競争軸そのものが変わった。 2024〜25年の「巨大な事前学習モデルを誰が作れるか」から、2026年には「推論時にどれだけ計算を動的配分できるか」「何時間・何日も道具を使って仕事を完遂できるか」「1Mトークン級コンテキストを実用速度で扱えるか」「FP4/MoE/スパースAttention/投機的デコードでtask-per-dollarを上げられるか」に移った。GPT-5.6 Solの`max/ultra`、Claude Opus 5のtest-time compute、Kimi K3のMXFP4量子化学習、DeepSeek V4のDSA、GLM-5.2/5.3系のIndexShare・MTP、Grok 4.6のagentic RLはすべてこの方向を示している。 [3]

第三に、中国の最大の競争力は「オープンウェイト＋巨大MoE＋効率化」の組み合わせである。 Kimi K3は2.8兆パラメータ、Qwen3.8系は2.4兆/95B active、DeepSeek V4 Proは1.6兆/49B activeである。Kimi K3は完全ウェイトを公開し、DeepSeekはMITライセンス、Qwenも2.4T-A95Bウェイトを公開した。ただしQwen3.8-Maxのクラウド版には、公開ウェイト版には含まれないVision、非thinking、標準1M context、組み込みtoolsなどがあるため、「Maxが完全にオープン」と表現するのは不正確である。 [4]

第四に、2026年後半〜2027年で最も確度が高いブレイクスルーは“次の巨大モデル”そのものではなく、モデル＋エージェント＋推論システムの統合である。 具体的には、①可変test-time compute、②階層型multi-agent、③長期記憶と1M超コンテキストの実用化、④FP4/FP8とスパースMoE、⑤投機的デコード、⑥モデル自身によるコード・カーネル・合成データ生成、⑦実環境での自己検証、⑧サイバー/生物能力の増大に伴う段階的アクセス制御が中心になる可能性が高い。これはすでにKimi K3の48時間チップ設計、Grok 4.6の長期agent RL、GLM-5.3の実世界専門家workflow post-training、GPT-5.6 Solのsubagent/Ultraなどに前兆が現れている。 [5]

最も端的に言えば、現在の構図は、

モデル知能：米国が小幅リード
オープンウェイト：明確に中国優位
最先端計算基盤：米国優位
効率化技術：ほぼ互角、一部中国優位
推論速度・商用クラウド統合：米国優位
エージェント能力：ほぼ同一世代
安全性評価・公開文書：米国優位

と整理するのが妥当である。 [6]

調査方法と比較上の注意点

今回の比較では、**企業公式リリース、公式モデルカード/API文書、公式ウェイトリポジトリを最優先**し、独立横断評価にはArtificial Analysisを主に使用した。企業自身のbenchmarkはモデルごとにharness、reasoning effort、tool使用条件が異なるため、異なるメーカーの数字をそのまま順位表として扱わず、「そのモデルが得意とする領域を確認する証拠」として用いた。特にFable 5のArtificial Analysis値はfallbackを含む構成であることに注意が必要である。 [7]

また、ユーザー指定の**MMLU、HELM、LLM-Scoreについて、今回の10モデルの公式発表には統一条件で比較できる横断値がほとんど存在しない**。2026年のリリース資料では、GPQA Diamond、Humanity's Last Exam、Terminal-Bench、DeepSWE、OSWorld、GDPVal-AA、Artificial Analysis Intelligence Indexなどが中心となっている。したがってMMLU/HELM/LLM-Scoreの数値は「未公開・統一比較不能」とし、存在しない値を補間しない方針とした。例えばGPT-5.6、Gemini 3.7 Flash、Kimi K3、DeepSeek V4 Proはいずれも、より難しいreasoning/agentic/実環境系評価を主要指標として公開している。 [8]

なお「パラメータ数未公開」は単なる情報不足ではなく、**米国クローズドモデルでは意図的な非公開が標準化**している。GPT-5.6、Claude 5系、Gemini 3.7、Grok 4.6、Muse Spark 1.2はいずれも公開ウェイトも正確なパラメータ数も公表していない。一方、中国勢はKimi、Qwen、DeepSeek、GLM系列でパラメータ数やMoE構造を公開する傾向が強い。したがって「パラメータ数」をそのまま能力比較に使うことはできないが、**技術透明性・自己ホスト可能性では大きな地域差がある**。 [9]

各フロンティアモデルの技術・商用比較

基本仕様・アーキテクチャ・学習・推論

モデル	公式リリース	公開仕様・Architecture	Training data / post-training	推論最適化・長文脈
GPT-5.6 Sol	OpenAI、**2026/7/9 GA**。6月末に限定preview。 [10]	パラメータ数・詳細architectureは**未公開**。1,050,000 context、最大128K出力、knowledge cutoff 2026/2/16。 [11]	完全なtraining corpusは未公開。OpenAIは「useful work per token」を重視した学習・post-trainingと、自身による production kernel最適化を説明。 [12]	reasoning effortはnone〜max。Ultraではsubagentsを追加。Cerebras版は最大約750 output tok/sとされる。 [13]

モデル	公式リリース	公開仕様・Architecture	Training data / post-training	推論最適化・長文脈
Claude Fable 5	Anthropic、**2026/6/9**。[14]	パラメータ数・内部architectureは**未公開**。1M context。Mythos 5と同一基盤の能力を、より厳しい安全機構付きで公開したモデル。[15]	corpus構成・token数は未公開。	adaptive reasoning / max effort。高リスクdomainではfallbackを含む安全構成。[16]
Claude Opus 5	Anthropic、**2026/7/24**。[17]	パラメータ数未公開、1M context、text/image→text。長期agentic codingとdeep reasoningを重点化。[18]	corpus非公開。post-trainingの詳細も限定的。	thinkingが基本、adaptive reasoningとhigh/max effortによりtest-time computeを拡大。[19]
Gemini 3.7 Flash	Google、**2026/8/13**。[20]	パラメータ数未公開。1M input、64K output。text/image/audio/video/PDF入力、native multimodal、search/function calling/computer-use対応。[21]	corpus詳細は未公開。	reasoning effort可変。Flash系として高throughputを重視。独立測定high設定で約340 tok/s。[22]
Grok 4.6	SpaceXAI/xAI、**2026/8/12**。[23]	パラメータ数・architecture未公開。500K context、text/image入力。[24]	4.5より長い追加training。**model-generated reasoning data、高品質engineering data、再生成SFT trajectories、agentic RL**を使用。[23]	agentic RLをcoding、kernel、web、CAD等に適用。high-effort reasoning。
Muse Spark 1.2	Meta Superintelligence Labs、**2026/8/5**。[25]	パラメータ数未公開。約1.049M context。text/image/audio/video入力、text出力。Muse Codeを同時展開。[26]	1.2固有のpretraining corpusは未公開。Spark 1.xはnative multimodal / agenticモデル系列。[27]	coding workflow、first-attempt accuracy、tool callingを重点強化。[28]

モデル	公式リリース	公開仕様・Architecture	Training data / post-training	推論最適化・長文脈
Kimi K3	Moonshot AI、**2026/7/14**。full weightsは7/27公開。 29	**2.8T total、約104B active、MoE**。896 experts中16を選択。Kimi Delta Attention、Attention Residuals、Stable LatentMoE、native vision、1M context。 30	corpus内訳は完全公開されていないが、refined data recipeとSFT以降のQATを説明。	**MXFP4 weights / MXFP8 activationsのQAT**、Quantile Balancing、Per-Head Muon、Gated MLA、vLLM向けKDA prefill cache。64 accelerator以上のsupernode推奨。 31
Qwen3.8-Max	Alibaba/Qwen、**2026/8/2発表、8/3 cloud GA**。 32	**2.4T total / 95B active MoE**。Cloud Maxは1M context、text/image/video、最大131K output。 33	詳細training corpusはlaunch資料では未公開。coding / cowork / agentic能力を重点化。	公開2.4T-A95BはFP8版も提供。ただし公開ウェイト版はMax cloudのvision、非thinking、標準1M、built-in toolsを完全には含まない。 34
DeepSeek-V4-Pro-0813	DeepSeek、**2026/8/13 GA**。V4 previewは4/24。 35	**1.6T total / 49B active MoE**、1M context。token-wise compression＋**DeepSeek Sparse Attention**。 36	完全なデータmixはlaunch資料では非開示。V4系列はagentic codingを明示的に強化。	DSpark speculative decoding、低精度cache等を利用。reasoning low/high/max。 37
GLM-5.3	Z.ai、**2026/8/14発表**。 38	5.3固有parameter countは未公表。**GLM-5.2と同一base**を使い、改善はpost-trainingのみ。GLM-5 lineageは744B total / 40B active。 39	GLM-5 baseは28.5T pretraining tokens。5.3はsenior engineerが数日かける実workflow、compute cluster、storage、内部docs/code repo等をpost-trainingへ導入。 40	基盤系列ではDSA、IndexShareにより1M context時per-token FLOPsを2.9×削減、MTP speculative decodingのacceptance lengthを最大20%改善。 41

ここで特に重要なのが、**中国勢は単に巨大化しているのではなくactive parameterを抑えている**点である。Kimi K3は2.8Tという非常に大きな総パラメータ数ながら16/896 expertsだけを選択し、DeepSeek V4 Proも1.6Tに対して49B active、Qwen3.8は2.4Tに対して95B activeである。これは「モデルを巨大化しつつ、一token当たりの実計算を抑える」設計であり、先端GPUへのアクセス制約が強い中国にとって戦略的に合理的である。 42

ベンチマーク、独立評価、商用化、ライセンス、安全性

モデル	主要評価	独立評価・実例	商用化 / License	Security / Safety
GPT-5.6 Sol	GPQA Diamond **94.6**、FrontierMath T1-3 **89.0**、OSWorld 2.0 **62.6**、BrowseComp **90.4**、Cyber CTF **96.7**、1M近傍long-contextでも高性能。[43]	AA Intelligence **61**。Model MLの業務評価ではPPT workflow完遂率100%などの事例。[44]	ChatGPT/API。closed proprietary。	能力上昇を理由に段階的preview。Preparedness Frameworkに基づく評価。SolはOpenAI基準のCyber Critical thresholdを越えていないと報告。[45]
Claude Fable 5	Anthropicがcoding、finance、physics等でfrontier級と報告。MMLU/HELMの統一公開値なし。[14]	AA Intelligence **62**。ただし評価構成にOpus fallbackを含む。[46]	Claude/API等。closed proprietary。標準価格はOpusより高いpremium tier。[47]	Mythos 5相当の能力にcyber/bio safeguardsを加え、高リスク時fallbackを使用。[48]
Claude Opus 5	長期coding/agent workflowを中心に改善。	AA Intelligence **63**で今回の対象中最高。1M context。[18]	Anthropic API、AWS、Google等。**$5 input/$25 output per MTok**。closed。[49]	Anthropic system card/RSPに基づく安全評価を公開。[50]
Gemini 3.7 Flash	DeepSWE **65.3**、Terminal-Bench 2.1 **85.8**、AutomationBench **30.4**、GDP.pdf **34.0**、HLE-Verified **53.6**、OSWorld2.0 **47.9**。[21]	AA Intelligence **56**、high設定で約340 tok/s。[51]	Gemini API/AI Studio/Vertex AI。導入価格$0.75/$3.75 MTok。closed。[20]	CBRN・cyber評価/mitigationを更新。[20]
Grok 4.6	GDPVal-AA v2 **1753**、DeepSWE1.1 **65.9**、CursorBench3.2 **69.9%**、FrontierCode **61.3**、Terminal-Bench3.0 **26%**。[23]	AA Intelligence **61**。	API/Cursor/Grok Build/OpenRouter/Vercel等。$2/$6 MTok。closed。[23]	過去最大規模のpre-deployment suiteと第三者post-deployment testingを実施したと説明。[23]

モデル	主要評価	独立評価・実例	商用化 / License	Security / Safety
Muse Spark 1.2	Googleの比較表ではDeepSWE **54.9**、Terminal-Bench2.1 **82.9**、GDPVal-AA **1628**。 21	AA Intelligence **57**。 52	Meta API/Muse Code。$1.25/$4.25 MTok前後。closed proprietary。 53	1.2固有safety cardの公開詳細は限定的。Muse系列はMeta Advanced AI Scaling Framework下でfrontier riskを評価。 27
Kimi K3	GPQA **93.5**、DeepSWE **67.5**、Terminal2.1 **88.3**、BrowseComp **91.2**、OSWorld Verified **84.8**、GDPVal Elo **1686**。 54	AA Intelligence **60**。今回の中国勢で最高、かつopen weight。 55	Kimi/Web/Work/Code/API。full weights公開。cache miss $3 input/$15 output。 31	独立した詳細safety system cardは確認できず。Custom Kimi K3 Licenseは法令順守を要求。大規模MaaSには追加契約条件あり。 56
Qwen3.8-Max	発表値：Terminal-Bench2.1 **86.6**、PaperBench **93.0**、SWE-bench Pro **67.7**、GPQA Diamond **92.6**、OSWorld-Verified **86.1**。 57	AA Intelligence **58**、Agentic系でも上位。 58	QwenCloud $2/$6 MTok。2.4T-A95B weights公開。ただしCloud Maxの全機能とは非同一。licenseは **qwen3.8-max custom**。 59	詳細なmodel-specific frontier safety cardは、今回確認した公開資料では限定的。中国国内サービスには生成AI規制・filing等が適用。 60
DeepSeek-V4-Pro-0813	HLE **42.7 / tools 60**、Terminal2.1 **87.9**、CyberGym **83.3**、DeepSWE **62.7**、Toolathlon **74.1**、Automation **31.8**。 61	AA Intelligence **53**。 62	app/web/API＋open weights。**MIT license**。 63	公式model-specific safety cardの公開は米国各社より薄い。service側のポリシーと中国規制が主要管理層。
GLM-5.3	Terminal-Bench3.0 **28.3**、DeepSWE **66.9**、Agents' Last Exam **28.5**、GDPVal-AA v2 **1769**、CyberGym **84.5**、ExploitBench **54.4**。 64	8/16時点でAAの5.3独立総合scoreはまだ確認できず。Reutersもcyber性能急上昇を報道。 65	GLM Coding Plan等で先行利用。weightsはsafety review後の公開予定で、8/16時点では5.3の完全公開ウェイトは未確認。GLM-5/5.2はMIT。 66	高度cyber機能はtrusted access、weight releaseも安全評価後という段階公開。 65

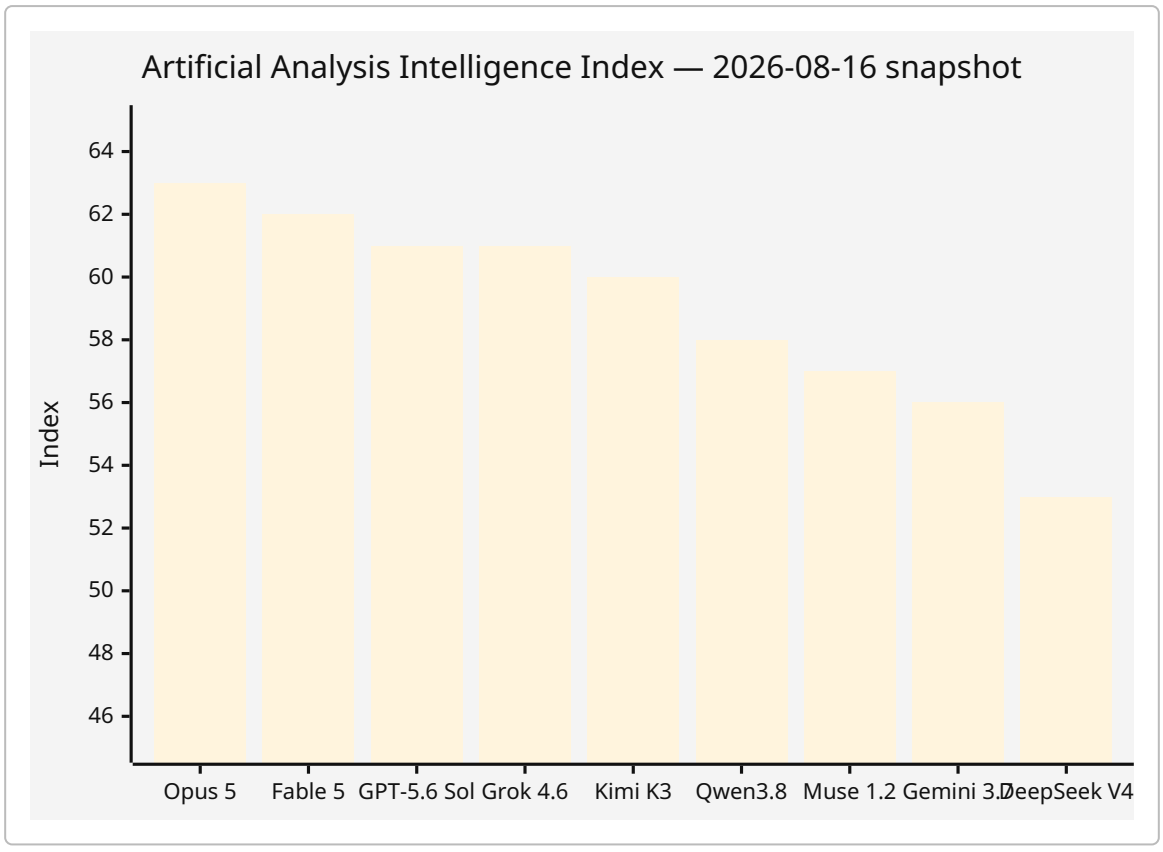
この表からもう一つ重要なことが分かる。**フロンティア性能と「重いモデル」は必ずしも同義ではない。** Gemini 3.7 FlashはAA 56ながらhigh設定で約340 tok/s、OpenAIはCerebras経由のGPT-5.6で最大約750 output tok/sを掲げる。一方Kimi K3はAA 60だがKimi公式APIに対するArtificial Analysis測定では約39.5 tok/

s、Qwen3.8 Maxは約47.1 tok/sである。したがって、現在の米中差の一部は「モデル知能」よりも**serving stack、hardware、interconnect、compiler、cache、provider implementation**に現れている。 67

また安全性に関しては、**米国勢の方がモデル固有のsystem card、capability threshold、CBRN/cyber evaluation、段階的アクセス制御を詳細に公開する傾向が強い**。中国側でもGLM-5.3のweight release延期やtrusted accessのような明確な変化が始まっているが、Kimi/Qwen/DeepSeekではモデル固有safety cardの情報量が比較的少ない。この差は「実際の安全性」そのものではなく、少なくとも**公開評価の透明性の差**として理解すべきである。 68

性能・リリース速度と「中国は半年遅れ」説の検証

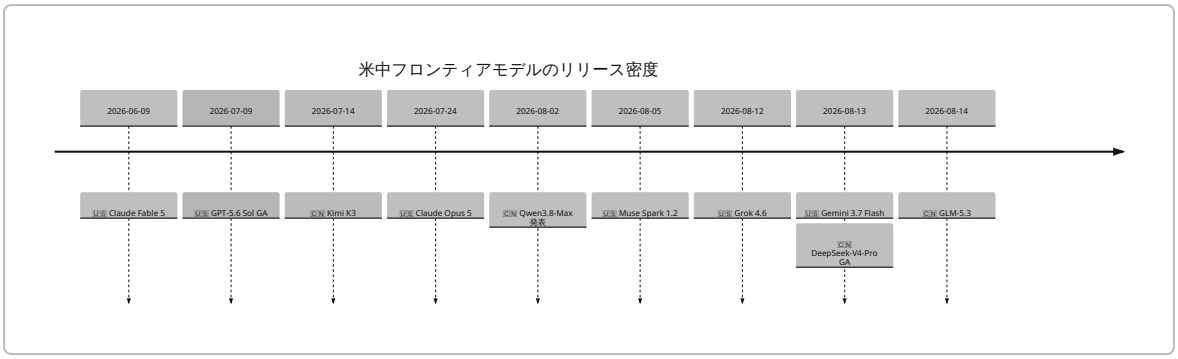
2026年8月16日時点で比較可能なArtificial Analysis Intelligence Indexを並べると次のようになる。effort設定はモデルごとにmax/xhigh/high等が異なり、Fable 5にはfallbackの影響があるため「絶対能力の厳密な順位」ではなく、**同一独立評価機関による現在地の近似**として読むべきである。Opus 5=63、Fable 5=62、GPT-5.6 Sol=61、Grok 4.6=61、Kimi K3=60、Qwen3.8 Max=58、Muse Spark 1.2=57、Gemini 3.7 Flash=56、DeepSeek V4 Pro=53である。 69 70



このチャートだけでも、**「米国と中国の間に一世代＝半年の空白がある」という説明は成立しにくい。** 中国最上位Kimi K3と米国最上位Opus 5の差は3ポイント、GPT-5.6 Sol/Grok 4.6との差は1ポイントである。さらにKimiはopen weightsである点を考えると、2025年までしばしば見られた「米国closed frontier → 数か月後に中国open modelが追う」という一方向の構造から、**同一リリース周期の中で競争する構造**へ移っている。 71

リリース日を並べても同じ傾向が見える。今回指定されたモデルだけを対象としたサンプルなので産業全体のrelease cadenceを意味するものではないが、中国側4社はKimi K3の7月14日からGLM-5.3の8月14日まで**約1**

か月に集中している。米国側もFable 5の6月9日からGemini 3.7 Flashの8月13日まで約2か月に集中しており、「米国が出したものを中国が半年後に模倣」という時系列ではない。 72



「半年遅れ」を要素別に分解すると、より正確な姿が見える。

競争軸	現在の判定	根拠
純粋な総合モデル能力	**米国が小幅リード、ほぼ同世代**	AA topは63対中国60。差は存在するが「半年」と呼べる規模ではない。 73
Coding / long-horizon agents	**ほぼ互角**	Kimi Terminal2.1 88.3、Qwen 86.6、DeepSeek 87.9。米国もGPT/Fable/Grokが各種agentic benchmark上位で、benchmarkごとに首位が入れ替わる。 74
Computer use / GUI agent	**中国に局所的リードあり**	Qwen発表のOSWorld-Verified 86.1、Kimi 84.8など、米国最上位と競争可能。 75
Open-weight frontier	**中国が明確に優位**	Kimi K3 full weights、DeepSeek V4 MIT、Qwen 2.4T-A95B weights。今回の米国対象6モデルはすべてclosed。 76
Inference efficiency algorithms	**互角〜中国が一部先行**	Kimi MXFP4/8 QAT、DeepSeek DSA、GLM IndexShare/MTP、Qwen massive sparse MoE。 77
商用推論速度 / full stack	**米国優位**	Gemini Flash約340 tok/s、GPT-5.6/Cerebras最大750 tok/s。一部中国flagshipは40〜50 tok/s級。 78
frontier hardware	**米国が依然かなり優位**	NVIDIA Rubinは2026 H2供給予定。一方、中国へのadvanced compute輸出にはlicense制約が残る。 79
国内AI siliconの代替能力	**中国が急速に改善**	HuaweiはAscend 950系を2026、960を2027へ展開するroadmapを示している。 80
safety transparency	**米国優位**	OpenAI/Anthropic/Google/xAIはモデル固有の能力・safety評価を多く公開。中国でもGLMなどに段階公開が出始めた。 81
release cadence	**実質互角**	今回の対象では中国4モデルが31日ほどに集中し、米国も2か月内に6モデル。半年の固定lagは観測されない。

競争軸	現在の判定	根拠
規制・データ環境	**優劣というより非対称**	米国は先端chip輸出を地政学的leverとして使い、中国はpublic generative-AI serviceへのfiling/content管理を強めている。training corpus自体は各社とも非公開が多く、単純な「米国はデータが多い」という定量結論は出せない。 82

ここから導ける結論は、**「半年遅れ説は、モデル能力については払拭された。一方、計算基盤・CUDA ecosystem・最先端半導体供給という構造的な米国優位は払拭されていない」**である。

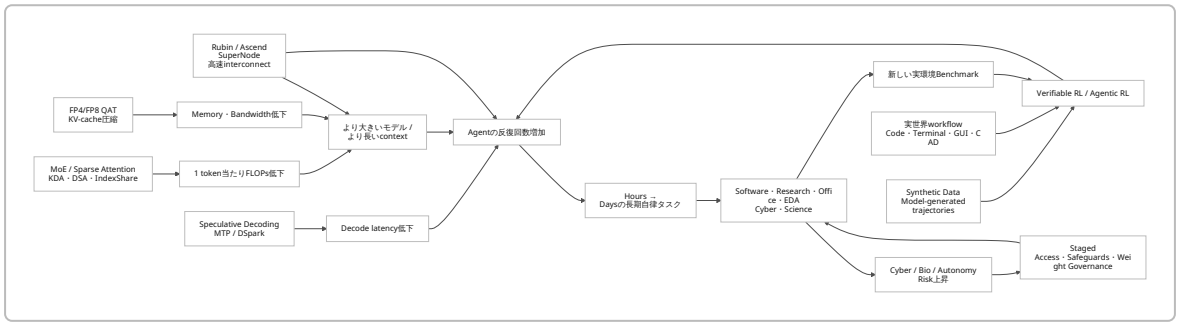
さらに、中国勢のキャッチアップの仕方には重要な特徴がある。米国と同じだけの先端GPUを確保することではなく、**MoE sparsity、低精度、sparse attention、speculative decoding、supernode、open-source serving stackを通じて、必要なGPU量を下げること**で差を埋めている。Kimi K3のMXFP4、DeepSeekのDSA、GLMのIndexShareはこの戦略の代表例である。 83

これは米国にとって無視できない。輸出規制が中国企業を完全にモデル競争から脱落させるのではなく、むしろ**計算効率を最優先するarchitecture研究を刺激する**副作用を持つ可能性がある、というのが現在のデータから得られる合理的な推論である。これは因果関係が公式に証明されたわけではないが、先端chip制約と、中国frontier modelで相次ぐsparsity/low-precision技術の併存からみて有力な解釈である。 84

次世代モデルを動かす技術ドライバー

2026年後半以降のブレイクスルーを考える際、最も重要なのは「GPT-6やKimi K4のparameter countが何兆になるか」ではない。すでに2〜3T級open MoEが出現しており、**総パラメータの追加だけでは性能・コスト・latencyの最適解にならなくなっている。** 次の競争はarchitecture、inference-time compute、agent system、data flywheel、hardware-software co-designを一体化する競争になる。 85

現在観測されている因果関係を整理すると以下のようになる。



この中で最も確度が高いのは**可変test-time computeの標準化**である。OpenAIはSolに`max`とmulti-subagentのUltra、AnthropicはOpus 5にadaptive reasoning/max、DeepSeekはlow/high/max、Kimiはmax thinkingを導入している。つまり「一つの固定モデルを一回forwardして答えを得る」形から、タスク価値に応じて**考えるtoken数、subagent数、verification回数、tool callsを動的に割り当てる**形へ、ほぼ全陣営が収束している。 86

第二のドライバーは**synthetic/verifiable data**である。Grok 4.6はmodel-generated reasoning dataと再生成SFT trajectoriesを明示し、GLM-5.3は実際のsenior engineer workflowをpost-trainingに取り込み、Kimi K3は自モデルのkernel最適化、compiler構築、チップ設計まで行っている。高品質なweb textを増やすだけ

ではなく、**モデル自身が仕事を実行し、結果をcompile/test/benchmarkして自動採点できる環境をtraining data generatorにする**段階へ移っている。[87]

第三は**attentionとmemoryの構造改革**である。Kimi KDA/AttnRes、DeepSeek DSA、GLM IndexShareはいずれもfull attentionをそのままスケールさせない方向である。DeepSeek V4を土台とした独立研究でも、必要なKV chunkだけをGPU上に保持し、平均physical KV cacheをfull-context baselineの13.5%まで圧縮しつつ精度を維持するアプローチが示されている。これは「context window 2M、10M」といった数字以上に、**必要な過去情報だけを検索・保持・再活性化するlearned memory**が重要になることを示唆する。[88]

第四はhardware-software co-designである。NVIDIA Rubinは2026年後半から市場投入され、NVIDIAはBlackwell比でMoE trainingの必要GPU数やtoken costを大きく削減できると説明している。一方Huaweiは Ascend 950を2026年、960を2027年に予定している。したがって2027年の競争は単純な「NVIDIA対Huawei」ではなく、**GPU/NPU、HBM、interconnect、precision format、compiler、kernel、MoE routing、serving frameworkを垂直統合したAI factory同士の競争**になる可能性が高い。[89]

2026年後半から2029年までのブレイクスルー・シナリオ

ユーザー指定の「短期6–12か月、中期12–24か月、長期24–36か月」を2026年8月16日起点で適用すると、短期は主に**2026年後半〜2027年前半**、中期は**2027年後半〜2028年前半**、長期は**2028年後半〜2029年前半**まで含む。したがって、中心分析は2026年後半〜2027年に置きつつ、後二者は指定された時間幅に合わせて先まで展望する。

Horizon	予想されるbreakthrough	主要driver	米中への影響	確度 / 最大の不確実性
短期 6–12か月	**Dynamic reasoning / test-time computeが標準化**。同一modelでfast〜deep〜multi-agentを連続制御。	GPT max/Ultra、Claude adaptive reasoning、DeepSeek low/high/max、Kimi maxの既存傾向。[86]	raw benchmarkよりtask completion/$が競争軸に。米国closed API、中国open/self-host双方で急進。	**高**。不確実性は推論compute増大をユーザーがどこまで支払うか。
短期	**数時間〜数日のcoding/research agentが実用品化**。resume、checkpoint、rollback、自己test、multi-agent delegationが標準機能へ。	Kimi 48時間chip-design、Grok long-running agent、Qwen長期coding、GPT/Claude agentic workflow。[90]	Software engineering、consulting、research、office automationの人間介入頻度低下。	**高**。最大問題は「長時間動くこと」と「最後まで正しいこと」の差。
短期	**FP4級training/inferenceと巨大Sparse MoEが主流化**。	Kimi MXFP4/8 QAT、Qwen 2.4T/95B active、Rubin FP4、DeepSeek sparse attention。[91]	中国はGPU制約を効率化で緩和。米国は同じ技術をRubin規模でさらに増幅。	**高**。accuracy degradationとcross-hardware portabilityが不確実。

Horizon	予想されるbreakthrough	主要driver	米中への影響	確度 / 最大の不確実性
短期	**1M contextから“persistent memory”へ**。context length自慢より、重要情報を選択・圧縮・再利用する能力へ。	KDA、DSA、IndexShare、KV cache圧縮研究。[92]	enterprise agentが巨大repo、mail、docs、DBを継続的に扱える。	**高〜中**。privacyとmemory poisoningが主要risk。
短期	**Open-weight Chinese frontierがclosed US frontierに継続的に数ポイント以内まで接近**。	Kimi60対米国61–63、Qwen58、DeepSeek53という現状と高速release cadence。[93]	「open=一世代遅れ」という前提が消失。企業がself-host vs APIを本格比較。	**高〜中**。米国勢が非公開の次世代architectureで再び大きく離す可能性。
中期 12–24か月	**Hierarchical multi-agent architecture**。単一LLMではなくplanner、coder、critic、researcher、memory、tool-routerのlearned orchestration。	OpenAI Ultra、Kimi 20+ subagents、Grok workflow fan-outなど現行systemの延長。[94]	有効computeをparameter scale以外で伸ばせる。企業業務のend-to-end automationが進展。	**中〜高**。agent間error propagationとcost制御。
中期	**Model-generated R&D data flywheel**。AIがkernel、compiler、benchmark、synthetic taskを作り、次世代modelを改善。	Grok generated trajectories、Kimi MiniTriton/kernel/chip、GPT production kernel optimization。[95]	AI研究開発速度そのものが加速。中国のcompute disadvantageをalgorithmic innovationで相殺する可能性。	**中〜高**。self-generated dataのmode collapseと評価汚染。
中期	**Multimodal action modelの統合**。vision/audio/video/computer-useから、continuous GUI/robotic/actionへ。	Gemini、Muse、Kimiのnative multimodalとGUI agent性能。[96]	Digital agentから一部physical automationへ範囲拡大。	**中**。robotics data、latency、安全certificationが制約。
中期	**評価革命**。静的QAから、1〜24時間以上の実務、cost、recovery、security、human oversightを測るbenchmarkへ。	Terminal-Bench、OSWorld、GDPVal、AutomationBench等への移行がすでに進行。[97]	「benchmark 95点」より「$10で8時間の仕事を完遂」の方が重要になる。	**高**。評価環境の再現性とcontaminationが課題。

Horizon	予想されるbreakthrough	主要driver	米中への影響	確度 / 最大の不確実性
中期	**Cyber capabilityの閾値問題が商用化を左右す**。	GPT-5.6 cyber評価、Grok safeguard拡張、GLM-5.3 ExploitBench急上昇とtrusted access。 [98]	top capabilityとpublic availabilityが分離し、「研究版」「trusted版」「public版」が一般化。	**高**。政府規制がどの閾値で介入するか。
長期 24–36か月	**Persistent learned memory / online adaptation**。巨大contextを毎回読み直さず、経験から内部/外部memoryを更新。	現在のsparse attention・KV圧縮・agent memoryの自然な延長。 [99]	個人・企業ごとに長期学習するagentが可能。	**中**。catastrophic forgetting、privacy、auditability。
長期	**Transformer+MoEの次の大規模architecture転換**。recurrent state、learned retrieval、conditional modules等が主役となる可能性。	KDA/AttnRes/DSAがすでにvanilla full-attentionから離れ始めている。 [100]	計算量の増え方が変われば、hardware gapを一時的にリセットし得る。	**低〜中**。現在のTransformer ecosystemの強さが非常に大きい。
長期	**限定domainでのclosed-loop AI researcher**。仮説→literature→code→experiment/simulation→分析→次の仮説を自律反復。	Kimiのscientific reproduction、compiler/chip実験、OpenAIのself-improvement benchmarkが前兆。 [101]	drug/materials、chip、software、math/scienceでR&D productivityが非線形に上昇する可能性。	**中〜低**。現実世界実験、novelty評価、safetyが律速。
長期	**広範なrecursive self-improvementによる急激な能力飛躍**	kernel/data/eval自動改善の延長として論理的には可能。	成立すれば米中差より「誰がclosed loopを最初に安定化するか」が重要になる。	**低**。現在の自己改善例はbounded engineering taskであり、一般的な自己再設計能力を証明していない。

特に2027年までの最有力シナリオは、**「単一の巨大LLMが突然AGIになる」のではなく、“モデルが自分で考える時間・subagent・memory・tools・verificationを配分できるAI operating system”へ進化すること**である。現在のfrontierモデルはすでにその入り口に立っており、Kimi K3が数十subagentで研究を実行し、OpenAI Ultraがsubagent computeを増やし、Grok Buildが多数agentをfan-outさせるという収斂が見られる。 [102]

この変化は経済性の評価も変える。たとえば、入力token単価が2倍でも、一度で正しくタスクを完了して再試行回数が半分になれば安い可能性がある。逆に安価なモデルでも100回tool callを繰り返せば高くなる。そのため2027年には**$/MTokより$/successful-task、tokens per completed workflow、human minutes saved**が重要な指標になると予想される。これはQwen、Kimi、Grok、Claude、GPTの評価がGDPVal、Terminal-Bench、OSWorld等へ移行していることとも整合する。 [103]

ハードウェア面では、短期的には米国優位がむしろ拡大する可能性がある。NVIDIA Rubinが2026年後半に立ち上がるため、米国系研究所は次世代training/inferenceで早期にその恩恵を受けられる。一方HuaweiもAscend 950→960と進むため、中国が停止するわけではない。したがって2027年には、**米国＝より強力なhardwareを大量投入、中国＝hardware制約をalgorithm/system efficiencyで圧縮**という二つの異なる最適化戦略がさらに鮮明になる可能性が高い。 [104]

安全保障面では、この技術競争が「どちらが最強modelを出すか」だけで終わらない点が重要である。米国は高度AI chip accessを外交・安全保障政策に組み込み、中国はopen-weightモデルを国際的ecosystem形成にも利用している。直近でも米国はAI・semiconductor supply chainをめぐり同盟国との連携を強化している。したがって2027年のAI競争は**model benchmark、cloud/API、open weights、chip supply chain、standards、安全性regimeを含む“AI stack圏”の競争**になる公算が大きい。 [105]

総合結論

2026年8月時点の最も重要な結論は、**米中AI競争を「米国が発明し、中国が半年後に追随する」という一次元モデルで理解する時代は終わった**ということである。現在は、

米国がclosed frontier、compute infrastructure、超高速serving、安全性評価で優位を保ちながら、中国がopen weights、巨大sparse MoE、low-precision、attention efficiency、cost-efficient self-hostingで先行または互角になる、非対称な競争へ変化している。 [106]

モデル性能だけを見るなら、「半年遅れ」はほぼ払拭された。Kimi K3のAA Intelligence 60に対し、GPT-5.6 Sol/Grok 4.6は61、Fable 5は62、Opus 5は63である。しかもKimi K3はfull-weight公開モデルである。Qwen3.8 Maxも58に達し、特定のcomputer-use、research、agentic benchmarkでは米国topモデルを上回る発表値を持つ。これはもはや異なる世代ではなく、**同一frontier band内の競争**である。 [107]

しかし、米国の最先端GPU・HBM・interconnect・CUDA・cloud ecosystemの優位は依然として重大であり、中国へのadvanced chip exportは制限され続けている。NVIDIA Rubinが2026年後半から普及すれば、その差が次のpretraining runで再び能力差として現れる可能性は十分ある。一方で、中国勢はすでにKimi K3のMXFP4、DeepSeekのDSA、GLMのIndexShare、巨大sparse MoEによって、そのhardware gapをモデル設計で吸収し始めている。したがって今後重要なのは、**「GPU数の差がmodel capabilityの差へどれだけ変換されるか」**である。 [108]

2026年後半〜2027年については、最も確度の高いbreakthroughは**巨大parameter jumpではなく、推論時compute scaling、multi-agent、persistent memory、FP4/MoE、speculative decoding、self-generated training trajectories、長期task verificationの統合**である。それによって、現在「チャットで賢い」モデルから、「数時間〜数日かかる仕事を、途中で失敗を発見・修正しながら最後まで完遂する」モデルへ競争軸が移る可能性が高い。 [109]

そしてこの段階では、米中の勝敗を単一benchmarkで決めることはますます難しくなる。**米国はcomputeとclosed-system integration、中国はopen ecosystemとefficiency innovationという異なる強みを持つため、2027年は一方が他方を恒常的に半年先行するより、architecture、agent、hardware、cost、安全性の各軸で首位が頻繁に入れ替わる市場になる**というのが、本調査から得られる最も蓋然性の高い見通しである。 [110]

1 9 18 69 73 110 Claude Opus 5 (max) - Intelligence, Performance & Price ...
https://artificialanalysis.ai/models/claude-opus-5?utm_source=chatgpt.com

2 79 89 104 108 NVIDIA Kicks Off the Next Generation of AI With Rubin
https://nvidianews.nvidia.com/news/rubin-platform-ai-supercomputer?utm_source=chatgpt.com

3 10 13 45 68 86 94 98 109 Previewing GPT-5.6 Sol: a next-generation model
https://openai.com/index/previewing-gpt-5-6-sol/?utm_source=chatgpt.com

4 5 29 30 31 42 77 83 85 90 91 92 100 101 102 Kimi K3 Tech Blog: Open Frontier Intelligence
https://www.kimi.com/blog/kimi-k3?utm_source=chatgpt.com

6 71 93 106 Kimi K3 (max) vs Claude Fable 5 (Adaptive Reasoning ...
https://artificialanalysis.ai/models/comparisons/kimi-k3-vs-claude-fable-5?utm_source=chatgpt.com

7 46 Claude Fable 5 (with fallback) - Intelligence, Performance ...
https://artificialanalysis.ai/models/claude-fable-5?utm_source=chatgpt.com

8 43 97 GPT-5.6: Frontier intelligence that scales with your ambition
https://openai.com/index/gpt-5-6/?utm_source=chatgpt.com

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

12 How GPT-5.6 fuses frontier intelligence with frontier efficiency
https://openai.com/index/gpt-5-6-frontier-intelligence-efficiency/?utm_source=chatgpt.com

14 72 Claude Fable 5 and Claude Mythos 5
https://www.anthropic.com/news/claude-fable-5-mythos-5?utm_source=chatgpt.com

15 Introducing Claude Fable 5 and Claude Mythos 5
https://platform.claude.com/docs/en/about-claude/models/introducing-claude-fable-5-and-claude-mythos-5?utm_source=chatgpt.com

16 48 Claude Mythos
https://www.anthropic.com/claude/mythos?utm_source=chatgpt.com

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

19 49 What's new in Claude Opus 5
https://docs.anthropic.com/en/docs/about-claude/models/whats-new-claude-4-8?utm_source=chatgpt.com

20 Gemini 3.7 Flash: our most intelligent workhorse model
https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-3-7-flash/

21 96 Gemini 3.7 Flash — Google DeepMind
https://deepmind.google/models/gemini/flash/

22 78 Gemini 3.7 Flash (high) API Provider Benchmarking & ...
https://artificialanalysis.ai/models/gemini-3-7-flash/providers?utm_source=chatgpt.com

23 87 95 Introducing Grok 4.6
https://x.ai/news/grok-4-6?utm_source=chatgpt.com

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

25 Meta AI Research: Muse Spark, Muse Glimmer and ...
https://ai.meta.com/research/?utm_source=chatgpt.com

26 52 53 Muse Spark 1.2 (xhigh) - Intelligence, Performance & Price ...
https://artificialanalysis.ai/models/muse-spark-1-2?utm_source=chatgpt.com

27 Introducing Muse Spark 1.1
https://ai.meta.com/blog/introducing-muse-spark-meta-model-api/?utm_source=chatgpt.com

28 Meet Muse Spark 1.2 and Muse Code
https://developer.meta.com/ai/resources/blog/build-with-muse-code/?utm_source=chatgpt.com

32 33 Qwen3.8-Max: A New Bar for Coding and Cowork
https://qwen.ai/blog?id=qwen3.8&utm_source=chatgpt.com

34 59 Qwen/Qwen3.8-2.4T-A95B
https://huggingface.co/Qwen/Qwen3.8-2.4T-A95B?utm_source=chatgpt.com

35 63 DeepSeek-V4-Pro GA Release
https://api-docs.deepseek.com/news/news260813?utm_source=chatgpt.com

36 88 DeepSeek V4 Preview Release
https://api-docs.deepseek.com/news/news260424/?utm_source=chatgpt.com

37 61 deepseek-ai/DeepSeek-V4-Pro-0813 · Hugging Face
https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro-0813

38 39 64 GLM-5.3 - Overview - Z.AI DEVELOPER DOCUMENT
https://docs.z.ai/guides/llm/glm-5.3

40 zai-org/GLM-5
https://huggingface.co/zai-org/GLM-5?utm_source=chatgpt.com

41 zai-org/GLM-5.2
https://huggingface.co/zai-org/GLM-5.2?utm_source=chatgpt.com

44 Claude Opus 5 (Adaptive Reasoning, Max Effort) vs GPT- ...
https://artificialanalysis.ai/models/comparisons/claude-opus-5-vs-gpt-5-6-sol?utm_source=chatgpt.com

47 Anthropic rolls out public version of Mythos without cybersecurity capability
https://www.reuters.com/technology/anthropic-rolls-out-public-version-mythos-without-cybersecurity-capability-2026-06-09/?utm_source=chatgpt.com

50 Model system cards
https://www.anthropic.com/system-cards?utm_source=chatgpt.com

51 Gemini 3.7 Flash (high) - Intelligence, Performance & Price ...
https://artificialanalysis.ai/models/gemini-3-7-flash?utm_source=chatgpt.com

54 74 GitHub - MoonshotAI/Kimi-K3: Open Frontier Intelligence · GitHub
https://github.com/MoonshotAI/Kimi-K3

55 107 Kimi K3 (max) - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/kimi-k3?utm_source=chatgpt.com

56 LICENSE · moonshotai/Kimi-K3 at main - Hugging Face
https://huggingface.co/moonshotai/Kimi-K3/blob/main/LICENSE?utm_source=chatgpt.com

57 Qwen3.8 Max Benchmarks & Speed (August 2026)
https://benchlm.ai/models/qwen3-8-max?utm_source=chatgpt.com

58 70 Qwen3.8 Max - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/qwen3-8-max?utm_source=chatgpt.com

60 China AI Regulation
https://www.deep-lex.com/ai-regulation-tracker/china?utm_source=chatgpt.com

62 Muse Spark 1.2 (xhigh) vs DeepSeek V4 Pro (Reasoning ...
https://artificialanalysis.ai/models/comparisons/muse-spark-1-2-vs-deepseek-v4-pro?utm_source=chatgpt.com

65 66 China's Z.ai says new model nears Anthropic's Mythos 5 in cyber-defence tests
https://www.reuters.com/technology/chinas-zai-says-new-model-nears-anthropics-mythos-5-cyber-defence-tests-2026-08-14/?utm_source=chatgpt.com

67 Previewing Ultrafast mode: GPT-5.6 Sol at up to 14X the ...
https://openai.com/index/previewing-ultrafast/?utm_source=chatgpt.com

75 Qwen3.8-Max: Features, Benchmarks, and Pricing
https://www.datacamp.com/blog/qwen3-8-max?utm_source=chatgpt.com

76 moonshotai/Kimi-K3 at c5d1dd4 - Initial commit - Hugging Face
https://huggingface.co/moonshotai/Kimi-K3/commit/c5d1dd4c428bd1ce8b88c5044f3b6ccde9e3b721?utm_source=chatgpt.com

80 China's Huawei hypes up chip and computing power plans ...
https://www.reuters.com/business/media-telecom/chinas-huawei-hypes-up-chip-computing-power-plans-fresh-challenge-nvidia-2025-09-18/?utm_source=chatgpt.com

81 GPT-5.6 System Card - Deployment Safety Hub - OpenAI
https://deploymentsafety.openai.com/gpt-5-6?utm_source=chatgpt.com

82 84 Revision to License Review Policy for Advanced ...
https://www.federalregister.gov/documents/2026/01/15/2026-00789/revision-to-license-review-policy-for-advanced-computing-commodities?utm_source=chatgpt.com

99 FlashMemory-DeepSeek-V4: Lightning Index Ultra-Long Context via Lookahead Sparse Attention
https://arxiv.org/abs/2606.09079?utm_source=chatgpt.com

103 GDPval-AA v2 Leaderboard
https://artificialanalysis.ai/evaluations/gdpval-aa?utm_source=chatgpt.com

105 US to tell partners they must pick sides in AI race with China
https://www.reuters.com/world/china/us-tell-partners-they-must-pick-sides-ai-race-with-china-2026-08-14/?utm_source=chatgpt.com