STRATEGY & TECHNOLOGY REPORT

中国製フロンティア LLM の現在地と日本企業の実装戦略

Kimi K3、GLM-5.2 (max)、Qwen3.7 Max、DeepSeek V4 Pro を中心に

GPT-5.6 Sol

米中性能差は消えたのか、追い越しは起きるのか。
公開ベンチマーク、独立評価、実装コスト、データガバナンスを分けて検証する。

調査基準日：2026 年 7 月 19 日（日本時間）

公開情報に基づく経営・調達・実装判断向けレポート

エグゼクティブ・サマリー

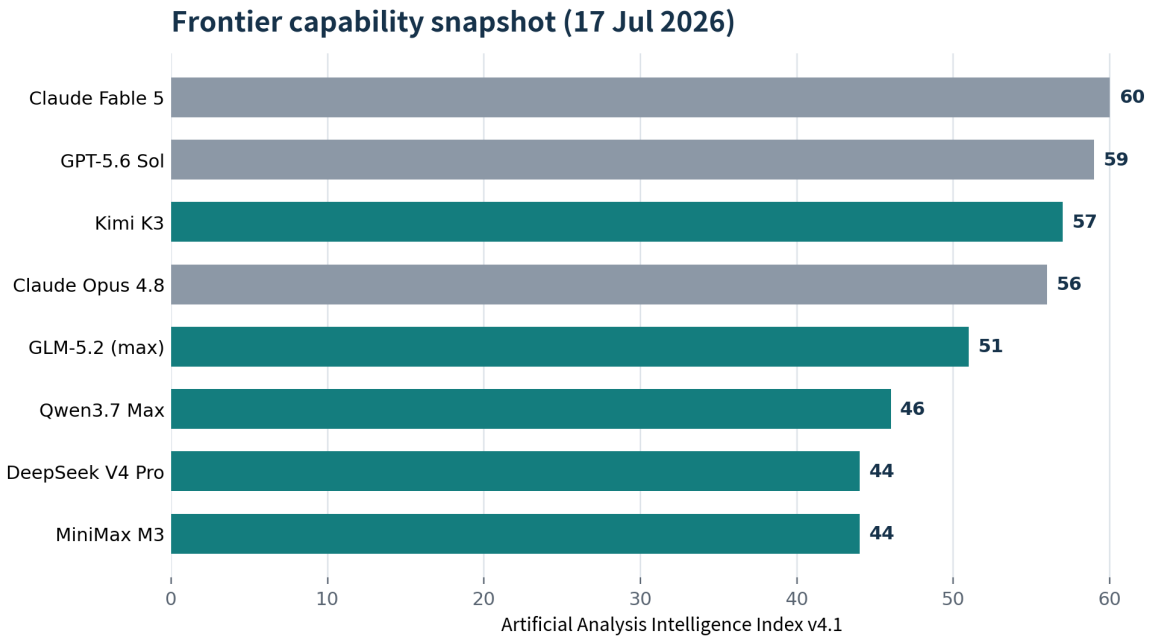
結論：中国製フロンティア LLM は、公開ベンチマークと一部の実務タスクでは米国系最上位と「ほぼ同じ競技場」に到達した。ただし、総合的・継続的に追い越したとはまだ言えない。日本企業は全面採用か全面排除かではなく、データ分類、提供経路、用途別評価を軸にした複線型ポートフォリオで使うべきである。

> 一言で言えば 「能力の追いつき」は概ね Yes、「企業で同じ安心感で使えるか」は条件付き、「全面的な追い越し」は未証明、ただしタスク別首位と価格・オープン性での先行はすでに起きている。

1. 性能差：Stanford AI Index 2026 は、米中のトップモデル差が 2026 年 3 月時点で 2.7% に縮小し、「実質的に閉じた」と評価した。一方、NIST/CAISI は DeepSeek V4 Pro を非公開課題で評価し、米国最先端に約 8 か月遅れるとした。両者は矛盾というより、対象モデル・時点・課題・評価方式が違う。[22][23][24]
2. 最新の位置：Kimi K3 は Artificial Analysis v4.1 で 57 点、GPT-5.6 Sol の 59 点、Claude Fable 5 の 60 点に接近した。WebDev Arena では暫定 1 位だが、Text Arena は 9 位タイ圏であり、モデル自身の公式説明も上位 2 モデルとの総合 UX 差を認める。[1][14][15][20][21]
3. 中国の先行領域：オープン重み、価格性能比、100 万トークン級の長文脈、フロントエンド実装、中国語・中国市場情報、効率化アーキテクチャで先行または首位を取る場面が増えた。Qwen 派生モデルが Hugging Face で 10 万超とされる普及ループも戦略的に重要である。[3][9][25][30]
4. 追い越し予測：2026〜2027 年は、単独の恒久王者ではなく、リリースごと・タスクごとに首位が入れ替わる「回転するフロンティア」が基本シナリオ。中国が持続的優位を作りやすいのはオープン実装・費用・産業導入、米国が優位を保ちやすいのは最上位の総合 UX、計算資源、クラウド・企業統制・安全評価の厚みである。[15][24][25]

5. 日本企業の行動：公開・低機密データで比較試験を直ちに始める一方、個人情報、営業秘密、未公開発明、M&A、重要インフラ・輸出管理対象は、承認のない中国向け消費者 SaaS へ入力しない。オープン重みの自社／専用基盤、契約済み地域型 API、DLP 付き AI ゲートウェイを使い分ける。[10][26][27][28][29]

図 1　独立評価によるフロンティア性能のスナップショット



出所：Artificial Analysis（2026-07-17 時点）[14][15][16][17][18]。推論努力、モデル世代、評価コストが異なるため、企業利用の等価性を示すものではない。

目次・読み方

> 更新性に関する注意 本市場は週単位で順位・価格・提供地域が変わる。特に Kimi K3 は調査基準日時点でモデル重みと完全な技術報告が未公開、Qwen3.8 Max はプレビュー段階である。採用前にモデル ID、スナップショット、契約条件、保存地域を再確認すること。

本報告では「中国製」を、主要な研究開発主体が中国に所在するモデルとして便宜的に括る。ただし企業リスクは国籍だけで決まらない。同じ DeepSeek V4 Pro でも、DeepSeek 直営 SaaS、東京リージョンの第三者 API、自社ネットワーク内のオープン重みでは、データ流通と統制可能性が大きく異なる。したがって、モデル、提供者、配備地域、契約、運用制御を分解して評価する。

1. 調査範囲と評価方法

対象は Kimi K3、GLM-5.2 (max)、Qwen3.7 Max、DeepSeek V4 Pro を中心に、MiniMax M3、ERNIE 5.1、Qwen のオープン系列等を補足した。基準日は 2026 年 7 月 19 日。モデル公開ブログ、公式モデルカード・API 文書、第三者評価、政府資料を突き合わせた。

証拠の優先順位

6. 一次情報：モデル仕様、ライセンス、価格、提供地域、制限事項。ベンダーの自己申告ベンチマークは事実として記録するが、優位性の結論には単独で用いない。
7. 共通の独立評価：Artificial Analysis の同一指数、Arena の人間選好、NIST/CAISI の非公開課題を重視する。
8. 企業実装：処理地域、学習利用、保存期間、監査、ID 管理、SLA、総タスク費用を性能と同格に扱う。
9. 日本適合：日本語、敬語、法務・知財語彙、国内固有の正確性は横断的な最新独立評価が不足するため、社内実課題で検証する。

> 重要な限界 ベンチマークは同じ名称でも、エージェント枠組み、最大ターン数、推論努力、コンテキスト管理、判定モデルが異なる。Stanford は広い市場スナップショット、NIST は特定モデルの非公開課題、Arena はユーザー選好を測る。単一の『何か月遅れ』へ還元しない。

2. 中国製 LLM は米国製に追いついたのか

2.1 結論を 4 つに分解する

問いの定義	判定	根拠	経営上の意味
公開総合ベンチマークで近いか	概ね Yes	トップ差 2.7%、Kimi K3 は独立指数で上位 3 点差。[15][23][24]	中国モデルを候補外にする合理性は薄れた。
特定タスクで首位か	すでに Yes	Kimi K3 が WebDev Arena 暫定 1 位。GLM-5.2 も長期コードで上位。[3][21]	用途別ルーティングが単一モデル契約より有利。
総合 UX・信頼性も同等か	まだ No	Kimi 自身が Fable 5／GPT-5.6 Sol との差を明記。NIST では DeepSeek に弱点。[1][22]	最終成果・高リスク判断は人間と別系統モデルで検証。
持続的な全面首位か	未証明	首位は数週単位で交代し、計算資源・投資・企業基盤は米国が厚い。[15][24][25]	ロックインを避け、交換可能な基盤を作る。

判定は本報告の統合評価。公開情報の変化に応じて四半期ごとに更新する。

2.2 「追いついた」と言える証拠

Stanford AI Index 2026 は、米中のトップモデルが 2025 年初から複数回首位を交換し、2026 年 3 月の米国トップのリードは 2.7%にすぎないとして、性能差が「実質的に閉じた」と整理した。さらに 2026 年 7 月、Artificial Analysis では Claude Fable 5 が 60、GPT-5.6 Sol が 59、Kimi K3 が 57 で、トップ 3 が 3 点幅に密集した。[15][23][24]

人間選好でも Kimi K3 は Text Arena で 1486 点、Gemini 3 Pro、GPT-5.6 Sol と同点圏に入り、WebDev Arena では 1679 点で Fable 5 の 1631、GPT-5.6 Sol の 1618 を上回った。ただし Kimi の票数は Text 3,024、WebDev 1,757 で暫定表示であり、長期安定順位ではない。[20][21]

価格・オープン性では追いつき以上の変化がある。GLM-5.2 と DeepSeek V4 Pro は MIT ライセンスの重みを公開し、DeepSeek は独立指数 44 点を非常に低いタスク費用で達成した。中国の戦略は、単一の最強モデルより、低価格・派生開発・導入データが次の改善を促す普及ループにある。[3][9][16][19][25]

2.3 それでも「全面的に同等」とは言えない理由

10. 非公開課題の差：NIST/CAISI は DeepSeek V4 Pro を H200/B200 と推奨設定で評価し、GPQA の自己申告値を再現できた一方、ARC-AGI-2、PortBench、サイバー課題では米国上位に遅れ、総合的に約 8 か月差と推定した。[22]

11. ギザギザの知能：数学でほぼ満点でも、サイバー、抽象推論、端末操作で弱いことがある。Stanford もエージェントが構造化課題の約 3 回に 1 回失敗し、一般ベンチマーク自体に 2〜42%の問題率があると指摘する。[23]

12. 製品成熟度：モデル品質だけでなく、権限管理、地域、保存、監査、事故対応、サポート、バージョン固定、コンテキスト互換性が必要。Kimi K3 は思考履歴に敏感で、曖昧な状況で過度に先回りする制限を自ら開示している。[1]

13. 日本語証拠の不足：英語・中国語中心の横断ベンチマークを、日本の契約、特許、金融、敬語、社内文書へそのまま外挿できない。JMMLU、JHumanEval、JBBQ 等を含む社内ベイクオフが必要である。[29][31]

3. 主要モデルの市場地図

表 1　技術仕様と提供形態（2026 年 7 月 19 日時点）

モデル	提供形態・ライセンス	規模	入力／文脈	企業上の要点
Kimi K3	API 提供中。重みは 7/27 までに公開予定、調査時点は未公開・ライセンス未確定。	2.8T、16/896 experts active	テキスト・画像／1M	独立指数 57、WebDev 暫定 1 位。非常に新しく、64 基以上のアクセラレータを推奨。[1][14]
GLM-5.2 (max)	オープン重み、MIT。Z.AI API／第三者／自社配備。	約 750B、40B active	テキスト／1M、最大出力 128K	長期コードと高速推論。大規模で自社配備の設備負担は重い。[3][4][16]
Qwen3.7 Max	独自モデル、重み非公開。Alibaba Cloud API。	非公表	テキスト／1M、最大出力 65,536	オフィス・長期エージェント。東京から利用可能でも Global scope は日本保存と同義でない。[5][6][7]
DeepSeek V4 Pro	オープン重み、MIT。直営 API／第三者／自社配備。	1.6T、49B active	テキスト／1M	極めて安価。非公開課題の弱点と直営サービスの中国保存に注意。[9][10][22]
MiniMax M3	オープン重み、MiniMax Community License。	約 428B、23B active	テキスト・画像・動画／1M	ネイティブマルチモーダル。商用条件と企業運用の成熟度を個別確認。[11][12]

注：T=trillion、B=billion、1M=100 万トークン。GLM の総パラメータは情報源により 744B／753B 表記があるため約 750B とした。「オープン重み」は学習データ・訓練コードまで完全公開する「オープンソース」と同義ではない。

表 2　独立評価・人間選好・概算価格

モデル	AA v4.1	Arena Text	Arena WebDev	API 価格概数（入力／出力、$/MTok）	読み方
Kimi K3	57	1486、9 位・暫定	1679、1 位・暫定	3.00／15.00（cache 0.30）	最上位に接近。出力が多く、トークン単価だけでなくタスク総額を見る。[1][14][20][21]
GLM-5.2 max	51	1468、30 位	1587、4 位	1.40／4.40（cache 0.26）	共通指数と WebDev は強いが、人間の一般選好は中位。非常に冗長。[16][20][21]
Qwen3.7 Max	46	1475、18 位・暫定	1516、17 位・暫定	約 1.48–2.50／4.42–7.50	高速・実務エージェント向け。価格は地域・スナップショットで差。[5][17][20][21]

モデル	AA v4.1	Arena Text	Arena WebDev	API 価格概数（入力／出力、$/MTok）	読み方
DeepSeek V4 Pro	44	1458、43 位	下位圏	0.43／0.87	価格性能比は突出。ただし AA で非常に冗長、NIST 非公開課題は弱い。[18][19][20][22]
MiniMax M3	44	—	1493、21 位	約 0.30–0.60／1.20–2.40	マルチモーダルと長文脈の代替候補。ライセンス精査が必要。[11][12][16][21]

出所：[14]〜[21]。Arena は 2026-07-16、Artificial Analysis は v4.1 の当時スナップショット。順位・価格は変動する。AA の旧版指数で Qwen3.7 Max が 56.6 と報じられた例があるが、ここでは現行 v4.1 の 46 に統一した。

4. モデル別の詳細評価

4.1 Kimi K3 — 最上位へ接近したが、まだ「発売直後」

> 位置づけ 今回の中国勢で最も強い総合性能。長期コーディング、知識労働、視覚入力で有力。ただし重み公開前、技術報告前、Arena 票数も少なく、本番標準への即時一本化は早い。

Moonshot AI は 2.8 兆パラメータ、1M 文脈、ネイティブ視覚を掲げ、Kimi Delta Attention、Attention Residuals、Stable LatentMoE を組み合わせ、896 expert のうち 16 を有効化する。K2 比で約 2.5 倍のスケーリング効率、MXFP4 重み／MXFP8 活性の量子化対応を説明する一方、本格配備には 64 基以上のアクセラレータを推奨する。[1]

独立評価では AA 57、GDPval-AA v2 1668、AA-Briefcase 1547 で、知識労働では GPT-5.6 Sol を上回る指標もある。しかし総合では Fable 5 と GPT-5.6 Sol に届かず、Kimi 自身も両者との UX 差を認める。思考履歴を完全に引き継がない枠組みで品質が不安定になり、曖昧な指示で予想外の判断をするという制限は、自律エージェントに権限を与える企業ほど重く見るべきである。[1][14][15]

14. 適合：公開情報リサーチ、視覚付き設計、フロントエンド、長期コード、資料作成のチャレンジャー。
15. 条件：重み・ライセンス・技術報告の公開後に供給網と自社配備コストを再評価。API では保存地域、DPA、保持、下請先を確認。

16. 避ける：ガードレールなしの自律購買、外部送信、コード投入、未公開発明の直営 SaaS 利用。

4.2 GLM-5.2 (max) — オープン重みの長期エンジニアリング

> 位置づけ MIT で企業が制御しやすく、長期コードと 1M 文脈の実用性が強み。総合指数 51 は中国オープン勢の上位だが、一般チャットの人間選好では順位が下がる。

GLM-5.2 は約 750B／40B active の MoE で、IndexShare により 4 つの疎注意層で索引器を共有し、1M 文脈でトークン当たり FLOPs を 2.9 倍削減、投機的デコードの受理長を最大 20%改善したとする。1M 入力、128K 出力、複数推論努力、Function Calling、構造化出力、MCP を備え、vLLM、SGLang、Ascend 等の配備経路がある。[3][4]

公式評価では Terminal-Bench 2.1 が 81.0、SWE-bench Pro が 62.1、FrontierSWE dominance が 74.4。独立 AA では 51、GDPval-AA v2 は 1524 で GPT-5.5 xhigh と同水準とされた一方、1 タスク当たり 43K 出力トークンと冗長で、Text Arena は 1468／30 位だった。これは『課題を完遂する強さ』と『会話として好まれること』の違いを示す。[3][16][20]

17. 適合：大規模コード監査、長期リファクタ、論文再現、社内専用基盤の候補。
18. 条件：巨大モデルの GPU、通信、運用人材、アップデート、脆弱性修正を TCO に含める。
19. 補足：Z.AI 直営 1M と、第三者クラウドでの実効コンテキスト上限は一致しないことがある。

4.3 Qwen3.7 Max — 独自モデルだが、企業 API とエージェント統合が強い

> 位置づけ 1M 文脈、ツール呼出し、オフィス自動化、非常に速い出力が魅力。一方、Max は重み非公開で、Qwen3 の Apache 2.0 オープン系列と混同してはならない。

Qwen3.7 Max はテキスト入力、1M 文脈、最大 65,536 出力を備え、OpenAI／Anthropic 互換 API、思考履歴の preserve_thinking、オフィス・コーディング・ロボティクス連携を訴求する。

公式事例では未知のプロセッサ向けカーネルを約 35 時間、1,158 回のツール呼出しで最適化した。これは持続的なエージェント能力の証拠候補だが、単一のベンダー実験でもある。[5][6]

独立 AA v4.1 は 46、約 200 tokens/s で高速。Text Arena は 1475／18 位、WebDev は 1516／17 位。総合首位ではないが、速度、企業 API、ツール統合を含む実効生産性で選ぶ価値がある。東京拠点の Model Studio で Qwen3.7 Max は Global scope として提供される一方、Japan scope に明示されるのは Qwen3.7 Plus である。『東京から呼べる』を『日本国内保存』と読み替えず、契約上確認する。[7][17][20][21]

20. 適合：オフィス文書、コード補助、長期エージェント、低待ち時間が重要な対話。
21. 条件：Max の重みは非公開。地域スコープ、処理・保存場所、下請先、スナップショット固定を契約化。
22. 代替：自社配備が要件なら Qwen3 の Apache 2.0 系列を別モデルとして再評価する。[30]

4.4 DeepSeek V4 Pro — 圧倒的な費用効率と、評価の食い違い

> 位置づけ MIT の 1.6T／49B active、1M 文脈。バッチ推論・コード・数学を安価に回す有力候補。ただし直営サービスのデータ取扱いと、非公開課題での性能差は明確な注意点。

DeepSeek V4 Pro は Compressed Sparse Attention と Heavily Compressed Attention を併用し、V3.2 比で 1M 文脈時の単一トークン FLOPs を 27%、KV cache を 10%に低減したとする。32T 超トークンで事前学習し、非思考、思考、Think Max を提供する。公式表では GPT／Claude に近い項目が並ぶが、これはベンダー選択の枠組み・課題である。[8][9]

NIST/CAISI は推奨設定と H200/B200 で自己申告の GPQA を再現したうえで、DeepSeek の自己申告上は 2 か月前の米国モデル級、非公開を含む CAISI 課題では 8 か月前の GPT-5 級と評価した。IRT 推定 Elo は DeepSeek 800、Opus 4.6 が 999、GPT-5.5 が 1260。数学は 96〜97%と強いが、ARC-AGI-2 46%、PortBench 44%、サイバー32%が弱かった。[22]

一方、費用効率は CAISI の 7 課題中 5 課題で米国の費用競争力ある参照モデルを上回り、AA でも 44 点を非常に低い費用で達成する。つまり『安い弱モデル』ではなく、大量の一次処理を担い、難問だけ上位モデルへ送るルータ設計に非常に向く。[18][19][22]

DeepSeek 直営のプライバシーポリシーは、入力・アップロード等を収集し、個人データを中国で直接収集・処理・保存すると明記する。モデル改善への利用からのオプトアウト権は示されるが、企業秘密を入力してよいことを意味しない。直営 SaaS と、自社配備した MIT 重みは別のリスクとして扱う。[10]

23. 適合：公開情報の大量要約、分類、コード候補、数学・技術一次解析、チャレンジャー／セカンドオピニオン。
24. 推奨経路：機密を扱うなら自社／専用基盤、または契約済みの地域型第三者提供。
25. 避ける：直営消費者サービスへの個人情報、未公開発明、顧客資料、ソースコードの投入。

5. その他の中国勢と競争構造

5.1 MiniMax、Baidu、Qwen オープン系列

MiniMax M3 は約 428B／23B active、1M 文脈、画像・動画を含むネイティブマルチモーダルで、MiniMax Sparse Attention により前世代比 9 倍の prefill、15 倍の decode、1M 時のトークン当たり計算量 1/20 を主張する。SWE-bench Pro 59、Terminal-Bench 2.1 66 という自己申告と、独立指数 44 が示すのは、最高総合性能よりマルチモーダルな自社配備の選択肢としての価値である。商用条件は Community License を精査する。[11][12][16]

Baidu ERNIE 5.1 は、ERNIE 5.0 から総パラメータを約 3 分の 1、active を約半分へ圧縮し、同規模モデルの約 6%の事前学習コストと説明する。検索、創作、エージェントに強いが、国際的な API 可用性と企業契約は用途・地域別の確認が必要である。[13]

Qwen では、Qwen3.7 Max は独自モデルである一方、Qwen3 の 235B、30B 等は Apache 2.0 で公開され、100 超の言語・方言を掲げる。USCC は Qwen 派生が Hugging Face で 10 万超と

し、オープンな派生・配備が改善と採用を循環させる中国の『第一のループ』を形成すると見る。[25][30]

5.2 競争の本体は「モデル 1 個」から「普及ループ」へ

中国の強みは、先端半導体の制約下で、疎な MoE、圧縮注意、低精度、推論最適化を高速に反復し、低価格とオープン重みで利用者を増やすことにある。利用が派生モデル、バグ修正、適用データを生み、製造、物流、ロボティクスの実世界データがさらに性能を押し上げる可能性がある。[1][3][9][11][25]

米国の強みは、2025 年の民間 AI 投資 2859 億ドル、中国 124 億ドルという 23 倍差、5,427 のデータセンター、複数の上位研究所、世界的クラウド・開発者基盤、企業向け統制の厚みにある。中国の政府系資金は民間投資統計に十分反映されない可能性はあるが、計算資源と資本の差は無視できない。[24]

6. 今後：中国は追い越すのか

> 基本シナリオ（2026〜2027 年） 総合首位は固定せず、米中複数社が数週〜数か月単位で入れ替わる。中国はオープン実装と費用、米国は最上位 UX・計算資源・企業基盤で相対優位を持つ。

表 3 「追い越し」のシナリオ別判断

観点	2026〜2027 年の見立て	確度	日本企業の備え
特定ベンチマーク／タスクで中国が首位	すでに発生。今後も頻発。	高	四半期でモデルを入れ替えられるルータと回帰試験。
オープン重みの性能・採用で中国が持続首位	最も起こりやすい。Kimi 公開後は特に注目。	高	重み評価、SBOM、ハッシュ、専用基盤の運用能力を確保。
価格性能比で中国が持続首位	DeepSeek 等で既に優位。米国の価格改定で変動。	高	単価ではなく正答 1 件／完遂 1 件当たり費用を計測。
総合能力で中国が 3 か月以上明確に首位	十分あり得るが未証明。米国の次世代投入で往復。	中	供給国を分けた二重化、API 抽象化、バージョン固定。
日本の高リスク企業利用で中国系が標準	性能以外の契約・信頼・地域・監査が制約。	低〜中	低リスクから始め、統制実績に応じて段階拡大。

「確度」は公開情報に基づく本報告の定性的シナリオ判断であり、統計的確率ではない。

6.1 中国が上回りやすい要因

26. オープン重みと低価格が、派生開発、世界的採用、実運用データ、次世代改善を循環させる。[25][30]
27. 先端 GPU 制約が、MoE 疎性、長文脈注意、FP4/FP8、キャッシュ、国内アクセラレータ対応を促す。[1][3][9][11]
28. 大きな中国語市場と製造・ロボティクスの産業基盤が、英語ベンチマーク外の実世界価値を生み得る。[24][25]
29. 企業間競争が激しく、Qwen、Moonshot、Z.AI、DeepSeek、MiniMax、Baidu 等が異なる強みで短周期に更新する。[13][15]

6.2 持続的な全面首位を阻む要因

30. 米国の計算資源、民間投資、最上位研究人材、クラウド・開発ツールの規模。[24]

31. モデルが巨大化し、オープン重みでも日本企業が現実的費用で自社配備できない場合がある。Kimi K3 の 64 基以上推奨は象徴的。[1]

32. 自己申告と非公開評価の差、ハルシネーション、過度な自律性、安全・サイバー評価の不足。[1][19][22][23]

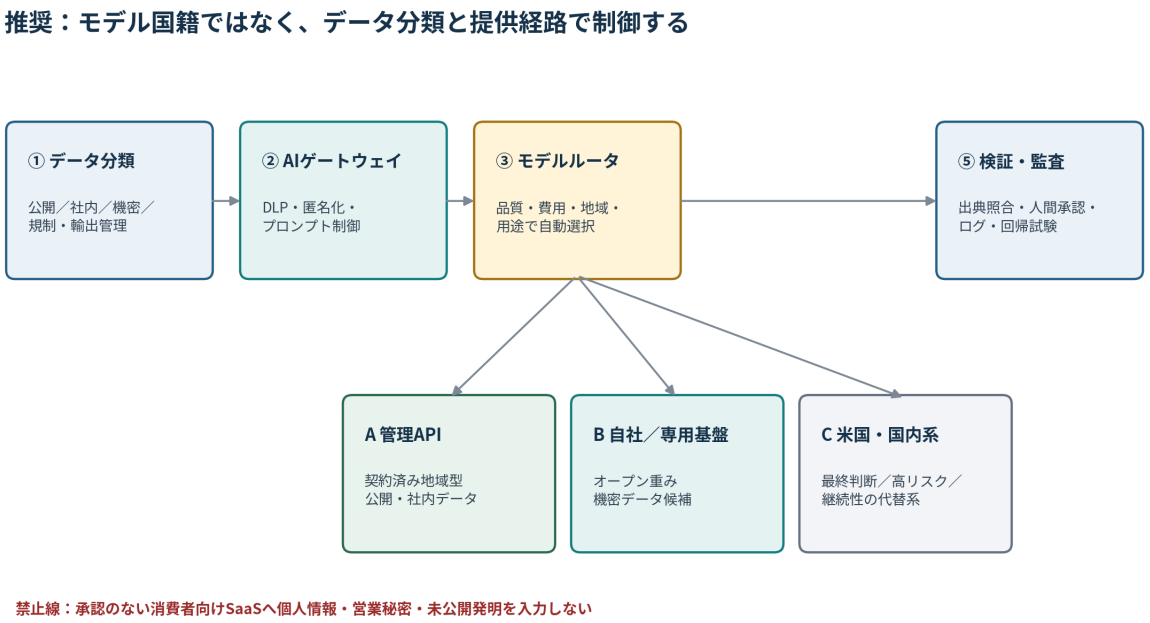
33. データ移転、検閲・拒否傾向、地政学、輸出規制、サービス継続、知財出所を含む企業信頼のコスト。

34. 日本語の法務・知財・顧客対応を保証する独立評価と、国内でのサポート・補償の厚みがまだ限定的。

7. 日本企業はどう使うべきか

> 推奨方針 中国製 LLM を『安い代替品』ではなく、高性能なチャレンジャー兼コスト最適化層として採用する。ただし国籍で一括判定せず、入力データと提供経路ごとに利用可否を決める。

図 2　日本企業向けの推奨制御アーキテクチャ



出所：本報告。データ分類→ゲートウェイ→用途別ルータ→検証・監査を共通化し、モデルを交換可能にする。

7.1 4 つの利用経路を混同しない

利用経路	使ってよい情報	主な利点	主なリスク／条件
① 直営の消費者 SaaS	公開情報・合成データのみ	最速で試せる。	規約、保存、学習、地域、管理機能が弱い。機密は不可。DeepSeek 直営は中国で処理・保存。[10]
② 契約済み企業 API	公開〜限定社内情報。契約と統制次第。	SLA、SSO、監査、地域選択の余地。	Global scope と国内保存を区別。DPA、保持、下請先、学習不使用を確認。[2][7][27]
③ 地域型第三者ホスティング	社内〜一部機密。審査済みの場合。	重みは中国、データ面は地域統制可能。	提供者の実装、ログ、モデル改変、供給網を再評価。

利用経路	使ってよい情報	主な利点	主なリスク／条件
④ 自社／専用・隔離基盤	機密候補。規制情報は個別承認。	最大のデータ制御、版固定、監査。	巨大 GPU、運用、脆弱性、重み真正性、安全評価、ライセンスの責任を自社が負う。

7.2 優先して試す用途

35. 公開情報の中国語リサーチ：規制、標準、特許、論文、競合、サプライヤー情報の抽出・翻訳・比較。
36. 高頻度の一次処理：分類、クラスタリング、要約、メタデータ付与、検索クエリ生成、コードテスト案。
37. 長文脈：大量の公開仕様書やリポジトリを読み、論点候補、矛盾、変更影響を抽出。
38. チャレンジャー：米国系・国内系モデルの回答と比較し、不一致を人間レビューへ回す。
39. コスト最適化：DeepSeek 等で一次処理し、難問だけ Kimi／GLM／米国最上位へエスカレーション。

7.3 人間承認を必須にする用途

40. 法務・知財：特許性、侵害、FTO、契約解釈、訴訟・規制判断。
41. 財務・経営：投資判断、開示、信用、価格、M&A、取締役会資料の最終結論。
42. 人事・医療・金融：個人に重大な影響を与える選別、助言、診断、与信。
43. 自律行動：送信、購入、削除、コード配備、権限変更、外部システム操作。
44. 顧客向け公開：誤り、偏り、権利侵害、ブランド毀損が直接発生する出力。

7.4 知財・研究開発部門への具体策

中国語の特許、公報、裁判・行政資料、技術コミュニティの一次資料を扱う場面では、中国モデルの語彙・検索発想が有利になり得る。ただし、未公開発明、クレームチャート、ライセン

ス交渉、発明者情報、M&A 対象、顧客秘密を直営 SaaS へ入れない。公開資料だけで候補を抽出し、必ず公的データベースの原文へ戻る。

45. 二段階検証：中国モデルが中国語資料から抽出・翻訳し、異なる系列のモデルと担当者が原文・書誌・引用箇所を検証。
46. 結論の分離：LLM は検索語、関連文献、相違点、反論候補を提示するが、新規性・進歩性・侵害・有効性の結論は専門家が行う。
47. 証跡：使用モデル ID、スナップショット、プロンプト、検索日時、引用 URL、原文抜粋、担当者承認を案件記録へ残す。
48. 発明漏えい対策：入力前に固有名詞、顧客名、未公開数値、図面、クレーム草案を除去し、必要なら隔離基盤のみで処理。
49. 反対意見の活用：米国系／国内系と中国系の結論が割れた案件を重点レビューし、探索範囲を広げる。

8. データ・法務・セキュリティの統制

日本の AI 事業者ガイドライン第 1.2 版は、人間中心、安全、公平、プライバシー、セキュリティ、透明性、アカウンタビリティ等をライフサイクルで扱う。経産省の契約チェックリストは、入力の学習利用、外部提供、外国にある第三者への個人データ提供、監査、ログ、規約改定を具体的に確認するよう促す。[26][27]

デジタル庁の政府向け第 2.0 版は、機密性 2 情報・個人情報が保存または学習される場合、重大影響業務、出力を人が判断せず用いる場合を高リスク軸として扱い、AI 統括責任者による継続的な監理を求める。民間企業への直接の法的義務ではないが、実務統制の有用な参照になる。[29]

> 法務上の注意　個人情報保護法上の外国第三者提供や委託該当性は、提供経路、契約、本人同意、相当措置等の事実関係で変わる。本報告は法律意見ではない。導入案件ごとに法務・プライバシー担当が確認する。

表 4　データ分類別の利用許容

区分	例	中国モデルの許容経路	必須統制
公開	公表済み Web、公開特許、公開 OSS、承認済み広報	直営 SaaS を含め利用可。ただし出力検証。	利用規約、引用、著作権、モデル ID、ログ。
社内	一般社内資料、非機微な手順、匿名化済みデータ	契約済み企業 API／地域型提供。直営消費者 SaaS は原則不可。	SSO/RBAC、DLP、学習不使用、保持・削除、監査。
機密	営業秘密、ソース、未公開 R&D、顧客資料、M&A	承認済み専用／自社基盤のみを原則。	隔離、暗号化、最小権限、出口制御、版固定、赤チーム。
規制・限定	個人・医療・金融、重要インフラ、防衛、輸出管理	外部 LLM は原則禁止。案件別の専用環境と法務承認。	法令・業法、データ所在、監査証跡、職務分離、停止機構。

8.1 調達・契約で最低限確認する 20 項目

50. 契約主体と準拠法、紛争解決、サービス提供国

51. 処理・保存・バックアップの国／リージョン

52. 再委託先・下請先と変更通知

53. 入力・出力・ログの学習利用と明示的な不使用

54. 保持期間、ゼロ保持の定義、削除証明

55. 個人データ、越境移転、DPA、監査協力

56. 暗号化、鍵管理、テナント分離

57. SSO、MFA、RBAC、退職者・委託先管理

58. 操作ログ、管理者ログ、エクスポート、保全期間

59. インシデント通知期限、調査協力、復旧

60. SLA、性能低下、利用停止、補償

61. モデル ID、スナップショット固定、変更予告

62. 価格改定、長文脈・推論トークン・キャッシュ費用

63. 入力と出力の権利、成果物利用、補償・免責

64. 学習データ・蒸留・コード出所に関する表明

65. 安全フィルタ、検閲・拒否傾向、地域差

66. オープン重みのライセンス、用途制限、再配布

67. 重みハッシュ、署名、SBOM、脆弱性対応

68. 事業継続、輸出規制・制裁・地政学による停止

69. 契約終了時の移行、データ返却、削除、代替モデル

9. 90 日で判断する導入ロードマップ

表 5　段階導入とゲート

期間	実施内容	成果物	次段階へのゲート
0〜15 日	データ区分、禁止線、対象 3 用途、承認者、候補経路を確定。	AI 利用基準、評価課題、契約質問票	個人・機密の扱いと責任者が明確。
16〜35 日	4〜6 モデルを同一の 100〜300 実課題で盲検比較。	品質・費用・速度・安全のスコアカード	現行モデル比の価値と弱点が再現可能。
36〜55 日	公開／合成データのサンドボックス。プロンプト注入と過度な自律性を試験。	赤チーム結果、失敗分類、停止条件	重大漏えい・無断操作の経路を閉じた。
56〜75 日	DLP 付きゲートウェイで限定社内データ。人間承認と二重検証。	監査ログ、運用手順、教育記録	法務・セキュリティ・業務責任者が承認。
76〜90 日	限定本番。モデルルータ、費用上限、回帰試験、代替系を有効化。	本番 KPI、SLA、切替・撤退計画	正答 1 件／完遂 1 件当たり費用と事故率が閾値内。

9.1 評価セットの推奨構成

70. 日本語の要約・起案・敬語・表現：30 件

71. 中国語資料・特許・規制の抽出と原文引用：25 件

72. 事実・引用・書誌の正確性：20 件

73. コード・ツール呼出し・長期実行：25 件

74. 長文脈の検索漏れ・矛盾検出：15 件

75. プロンプト注入、権限逸脱、データ流出：20 件

76. 地政学・歴史・規制テーマの拒否／偏り：10 件

77. 待ち時間、総トークン、再試行、タスク総費用：全件

9.2 合格指標

単純な正解率だけでなく、タスク完遂率、引用適合率、根拠の再現性、重大誤り率、無断ツール実行率、データ漏えい率、拒否の過不足、日本語レビュー時間、総コスト、モデル更新後の回帰率を測る。トークン単価ではなく「承認可能な成果 1 件当たり費用」を最終指標にする。

> 停止条件の例 機密送信、無承認の外部操作、引用の捏造、規制対象データの越境、重大判断の自動確定、モデル変更後の重大回帰が 1 件でも発生した場合は自動停止し、原因と統制を再審査する。

10. 推奨ポートフォリオと意思決定

表 6　用途別の第一候補・代替・制約

用途	第一候補	代替／検証系	導入条件
最高性能の公開情報リサーチ・資料作成	Kimi K3 をパイロット	米国最上位＋担当者	重み・技術報告の公開待ち。本番は出典検証と権限制限。
長期コード・大規模リポジトリ	GLM-5.2 max	Kimi K3／米国コードモデル	社内基準、テスト、変更範囲、コミット権限を明示。
高速なオフィス・ツール連携	Qwen3.7 Max	Qwen Plus／米国・国内系	処理地域、スナップショット、DPA、総費用を確認。
大量の安価な一次処理	DeepSeek V4 Pro	MiniMax M3／小型 Qwen	公開・匿名化、または自社／地域型提供。難問を上位へ転送。
機密の自社配備	GLM／DeepSeek／MiniMax を比較	国内系オープン重み	GPU TCO、重み真正性、ライセンス、安全・日本語評価。
高リスクの最終判断	人間＋実績ある企業基盤	中国モデルは反対意見・探索	自動確定を禁止し、一次資料と二重承認を必須化。

10.1 取締役会／経営会議で決める 5 項目

78. 中国モデルを候補に含める。ただし直接 SaaS ではなく、承認済み提供経路を原則とする。

79. 公開・低機密の 3 用途で 90 日評価を開始し、現行米国系／国内系と盲検比較する。

80. AI ゲートウェイ、モデルルータ、監査ログへ投資し、単一ベンダーロックインを避ける。

81. 営業秘密、未公開発明、個人・規制情報の禁止線と例外承認者を明文化する。

82. 四半期ごとにモデル・価格・地域・規約・安全評価を更新し、撤退・切替試験を行う。

> 最終提言 日本企業にとって最悪の選択は、中国モデルを無条件に避けて費用・性能・知見を失うことでも、安さだけで機密を直営サービスへ送ることでもない。最善は、データを守る共通基盤の上で、中国・米国・国内モデルを競わせ、用途別に勝者を変えられる運用能力を自社に残すことである。

付録 A　評価スコアカード

評価軸	測定例	重みの目安	最低条件
業務品質	完遂率、正確性、網羅性、指示遵守	30%	現行基準を上回る、またはレビュー時間を有意に短縮。
根拠・再現性	引用適合、一次資料到達、同一版の再現	15%	重要主張に検証可能な根拠。引用捏造ゼロ。
日本語適合	敬語、法務・知財語彙、読みやすさ	10%	専門担当者が許容する誤り率。
安全・権限	注入耐性、無断操作、漏えい、停止	20%	重大違反ゼロ、操作は最小権限と承認付き。
データ・契約	地域、学習、保持、監査、権利	15%	データ区分に合う契約と技術統制。
経済性・運用	成果 1 件費用、速度、可用性、切替	10%	予算内、SLA、代替モデルへ切替可能。

重みは用途で調整する。知財・法務では根拠・再現性とデータ統制を増やし、バッチ分類では経済性を増やす。平均点だけで合格させず、安全・データ・引用捏造に独立した失格条件を置く。

失敗分類

83. F1 事実誤り：存在しない仕様、価格、判例、特許、公報、引用。

84. F2 推論誤り：根拠は正しいが結論が飛躍、反証を無視。

85. F3 指示逸脱：範囲外変更、許可のないツール・送信・削除。

86. F4 データ統制：禁止情報の送信、ログ欠落、地域・保持違反。

87. F5 日本語品質：敬語、固有名詞、専門語、ニュアンスの誤り。

88. F6 安全・偏り：過剰拒否、無警告の危険助言、政治・歴史テーマの偏り。

89. F7 運用：タイムアウト、文脈破綻、版変更、費用暴走。

付録 B　用語と読み違いを防ぐポイント

オープン重み：学習済み重みを取得できること。学習データ、訓練コード、評価データまで公開されるとは限らない。

オープンソース：本来はソースコードとライセンスの概念。LLM では曖昧に使われるため、本報告は原則「オープン重み」と記す。

MoE / active parameters：多数の expert の一部だけをトークンごとに使う設計。総パラメータが大きくても推論計算量は active 側に近づくが、保存・通信は重い。

1M context：100 万トークンの入力枠。100 万を入れて重要情報を確実に使えること、低遅延・低費用で使えることとは別。

reasoning effort / max：推論トークンと探索量を増やす設定。精度が上がる場合がある一方、費用・時間・冗長性が増える。

Arena：匿名のモデル出力を人が比較する選好評価。会話の好ましさを反映するが、票数、スタイル、利用者層の影響を受ける。

Artificial Analysis Index：複数の知識・推論・コード・エージェント課題を統合した独立指数。版ごとに構成が変わるため、異なる版の点数を直結しない。

データ所在：API の URL や利用地点ではなく、推論、ログ、バックアップ、サポートアクセスが実際にどこで行われるか。契約で確認する。

参考文献

本文中の [n] は以下に対応する。Web 情報はすべて 2026 年 7 月 19 日に最終確認。ベンチマーク、価格、提供地域、規約は更新されるため、導入時に原典を再確認する。

[1] Moonshot AI (2026-07-16), “Kimi K3: Open Frontier Intelligence.” https://www.kimi.com/blog/kimi-k3（最終確認：2026-07-19）

[2] Moonshot AI, “Kimi Business: Enterprise AI Without Enterprise Complexity.” https://www.kimi.com/business（最終確認：2026-07-19）

[3] Z.AI, “GLM-5.2 Model Card,” Hugging Face. https://huggingface.co/zai-org/GLM-5.2（最終確認：2026-07-19）

[4] Z.AI Developer Documentation, “GLM-5.2 — Overview.” https://docs.z.ai/guides/llm/glm-5.2（最終確認：2026-07-19）

[5] Qwen Team (2026-05), “Qwen3.7: The Agent Frontier.” https://www.alibabacloud.com/blog/qwen3-7-the-agent-frontier_603154（最終確認：2026-07-19）

[6] Alibaba Cloud Model Studio, “Text generation models.” https://help.aliyun.com/en/model-studio/text-generation-model（最終確認：2026-07-19）

[7] Alibaba Cloud Model Studio, “Qwen Context Cache” (deployment scopes included). https://help.aliyun.com/en/model-studio/context-cache（最終確認：2026-07-19）

[8] DeepSeek (2026-04-24), “DeepSeek V4 Preview Release.” https://api-docs.deepseek.com/news/news260424/（最終確認：2026-07-19）

[9] DeepSeek AI, “DeepSeek-V4-Pro Model Card,” Hugging Face. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro（最終確認：2026-07-19）

[10] DeepSeek (updated 2026-02-10), “DeepSeek Privacy Policy.” https://cdn.deepseek.com/policies/en-US/deepseek-privacy-policy.html（最終確認：2026-07-19）

[11] MiniMax (2026-06), “MiniMax M3: Frontier Coding, 1M Context, Native Multimodality.” https://www.minimax.io/blog/minimax-m3（最終確認：2026-07-19）

[12] MiniMax AI, “MiniMax-M3 Model Card,” Hugging Face. https://huggingface.co/MiniMaxAI/MiniMax-M3（最終確認：2026-07-19）

[13] Baidu (2026-05-09), “ERNIE 5.1 Officially Released!” https://ernie.baidu.com/blog/posts/ernie-5.1-0508-release/（最終確認：2026-07-19）

[14] Artificial Analysis (2026-07-17), “Kimi K3 achieves #3 in the Artificial Analysis Intelligence Index.” https://artificialanalysis.ai/articles/kimi-k3-achieves-3-in-the-artificial-analysis-intelligence-index-comparable-to-opus-4-8-and-gpt-5-5（最終確認：2026-07-19）

[15] Artificial Analysis (2026-07-17), “Four frontier launches in eight days.” https://artificialanalysis.ai/articles/four-frontier-launches-in-eight-days-six-labs-now-field-a-model-above-50-on-the-artificial-analysis-intelligence-index（最終確認：2026-07-19）

[16] Artificial Analysis (2026-06-16), “GLM-5.2 is the new leading open weights model.” https://artificialanalysis.ai/articles/glm-5-2-is-the-new-leading-open-weights-model-on-the-artificial-analysis-intelligence-index（最終確認：2026-07-19）

[17] Artificial Analysis, “Qwen3.7 Max — Intelligence, Performance & Price Analysis.” https://artificialanalysis.ai/models/qwen3-7-max（最終確認：2026-07-19）

[18] Artificial Analysis, “DeepSeek V4 Pro — Intelligence, Performance & Price Analysis.” https://artificialanalysis.ai/models/deepseek-v4-pro（最終確認：2026-07-19）

[19] Artificial Analysis (2026-06-15), “Artificial Analysis Intelligence Index v4.1.” https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v4-1（最終確認：2026-07-19）

[20] Arena (2026-07-16 snapshot), “Text Arena Overall.” https://arena.ai/leaderboard/text（最終確認：2026-07-19）

[21] Arena (2026-07-16 snapshot), “Code Arena | WebDev Overall.” https://arena.ai/leaderboard/code（最終確認：2026-07-19）

[22] NIST CAISI (2026-05), “CAISI Evaluation of DeepSeek V4 Pro.” https://www.nist.gov/news-events/news/2026/05/caisi-evaluation-deepseek-v4-pro（最終確認：2026-07-19）

[23] Stanford HAI (2026), “2026 AI Index Report — Technical Performance.” https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance（最終確認：2026-07-19）

[24] Stanford HAI (2026), “The 2026 AI Index Report.” https://hai.stanford.edu/ai-index/2026-ai-index-report（最終確認：2026-07-19）

[25] U.S.-China Economic and Security Review Commission (2026-03-23), “Two Loops: How China’s Open AI Strategy Reinforces Its Industrial Dominance.” https://www.uscc.gov/sites/default/files/2026-03/Two_Loops--How_Chinas_Open_AI_Strategy_Reinforces_Its_Industrial_Dominance.pdf（最終確認：2026-07-19）

[26] 経済産業省・総務省（2026）「AI 事業者ガイドライン（第 1.2 版）」 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20260331_report.html（最終確認：2026-07-19）

[27] 経済産業省（2025）「AI の利用・開発に関する契約チェックリスト」 https://www.meti.go.jp/policy/mono_info_service/connected_industries/sharing_and_utilization/20250218003-ar.pdf（最終確認：2026-07-19）

[28] 個人情報保護委員会（2023）「生成 AI サービスの利用に関する注意喚起等について」 https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/（最終確認：2026-07-19）

[29] デジタル庁（2026-06-12）「行政の進化と革新のための生成 AI の調達・利活用に係るガイドライン 第 2.0 版」 https://www.digital.go.jp/assets/contents/node/information/field_ref_resources/decb64eb-f26e-41cb-8d37-f3dd173108b8/dc8912c3/20260612_resources_standard_guidelines_guideline_07.pdf（最終確認：2026-07-19）

[30] Qwen Team (2025-04-29), “Qwen3: Think Deeper, Act Faster.” https://qwenlm.github.io/blog/qwen3/（最終確認：2026-07-19）

[31] LLM-jp「日本語 LLM まとめ」（日本語評価ベンチマーク一覧を含む） https://llm-jp.github.io/awesome-japanese-llm/（最終確認：2026-07-19）

[32] Reuters (2026-07-17), “China’s Moonshot unveils world’s largest open AI model, closing in on US rivals.” https://www.reuters.com/world/china/chinas-moonshot-unveils-worlds-largest-open-ai-model-closing-us-rivals-2026-07-17/（最終確認：2026-07-19）

調査上の留保

90. Kimi K3 の重みは 2026 年 7 月 27 日までに公開予定で、調査基準日時点では未公開。公開後のライセンス、重み一致、第三者再現、量子化品質で評価が変わり得る。

91. モデル名の max、thinking、high 等は推論努力を含む。非思考版や異なる提供者との点数比較は同一条件ではない。

92. 価格は米ドル建ての 100 万トークン当たり概数。キャッシュ、長文、画像、ツール、推論トークン、地域、契約割引、再試行、サポートを含む総額ではない。

93. 日本語・日本法務・日本の知財実務を対象に、4 モデルを同一条件で比較した十分な最新独立データは確認できなかった。

94. 法務・規制・情報セキュリティに関する記載は一般的な実務整理であり、個別案件への法律意見ではない。

> 再評価日 Kimi K3 の重み・技術報告公開後（2026 年 7 月末）に速報更新し、遅くとも 2026 年 10 月に価格、Arena 票数、独立指数、提供地域、契約条件を再評価する。