

[CLASSIFICATION: OPEN-WEIGHT] | [SUBJECT: KIMI K3] | [DATE: 2026-08-01]

Kimi K3 調査報告：オープンウ エイットの衝撃と実態

事実関係・技術仕様・第三者評価・ライセンス及び政策的論点の完
全解剖

SOURCE: CLAUDE OPUS 5

INTELLIGENCE DOSSIER

エグゼクティブ・シンセシス：3つの重要論点



フロンティア級、 ただし絶対王者ではない

- 総パラメータ2.8兆のMoEアーキテクチャ。
- オープンウェイトとして最高峰の知能（Intelligence Index 57点）を示すが、総合首位には届かず。フロントエンド・コード生成に特筆すべき強みを持つ。



米中ハイテク摩擦の 新たな火種

- 米国政府によるAnthropic「Fable」の蒸留疑惑および輸出規制違反の指摘。
- 現時点で証拠は未提示だが、エンティティ・リスト入りのリスクなど、採用にあたっての地政学的監視が必須。



「オープン」の幻想と ライセンスの罠

- 総容量1.56TB。最低でも「8×H100」を要求し、一般的なローカル環境では実行不可能。
- 完全なオープンソース（MIT等）ではなく、商用しきい値を持つ独自の「Kimi K3ライセンス」が適用される。

[SUBJECT: TIMELINE VERIFICATION] | [MYTH VS FACT]

タイムライン検証：情報とノイズの分離

市場のノイズ

~~2026年7月26日夜：1日前倒し
でサプライズ公開された。（※具
体的時刻は未確認の二次情報）~~

~~2026年7月28日：オープンウェイトが
一般公開された。（※不正確。28日は
サーバー対応発表や米中の応酬の日付）~~

7月16日

7月26日

7月28日

7月16日

7月26～27日

7月27日

検証済み 事実

API・製品（Kimi.com, Kimi Code）として公開。上海WAIC開催と同期。

Moonshotの事前予告通り、Hugging Face上で完全なウェイト（約1.56TB）が公開される。

技術報告書（arXiv:2607.24653）およびインフラ技術（MoonEP, FlashKDA等）の公開。

アーキテクチャ・ブループリント：2.8Tパラメータの実態

スケール

- Mixture-of-Experts (MoE)
- 総パラメータ：2.8兆 (2.78T)
- 1トークンあたりの活性パラメータ：104B

極疎ルーティング (Stable LatentMoE)

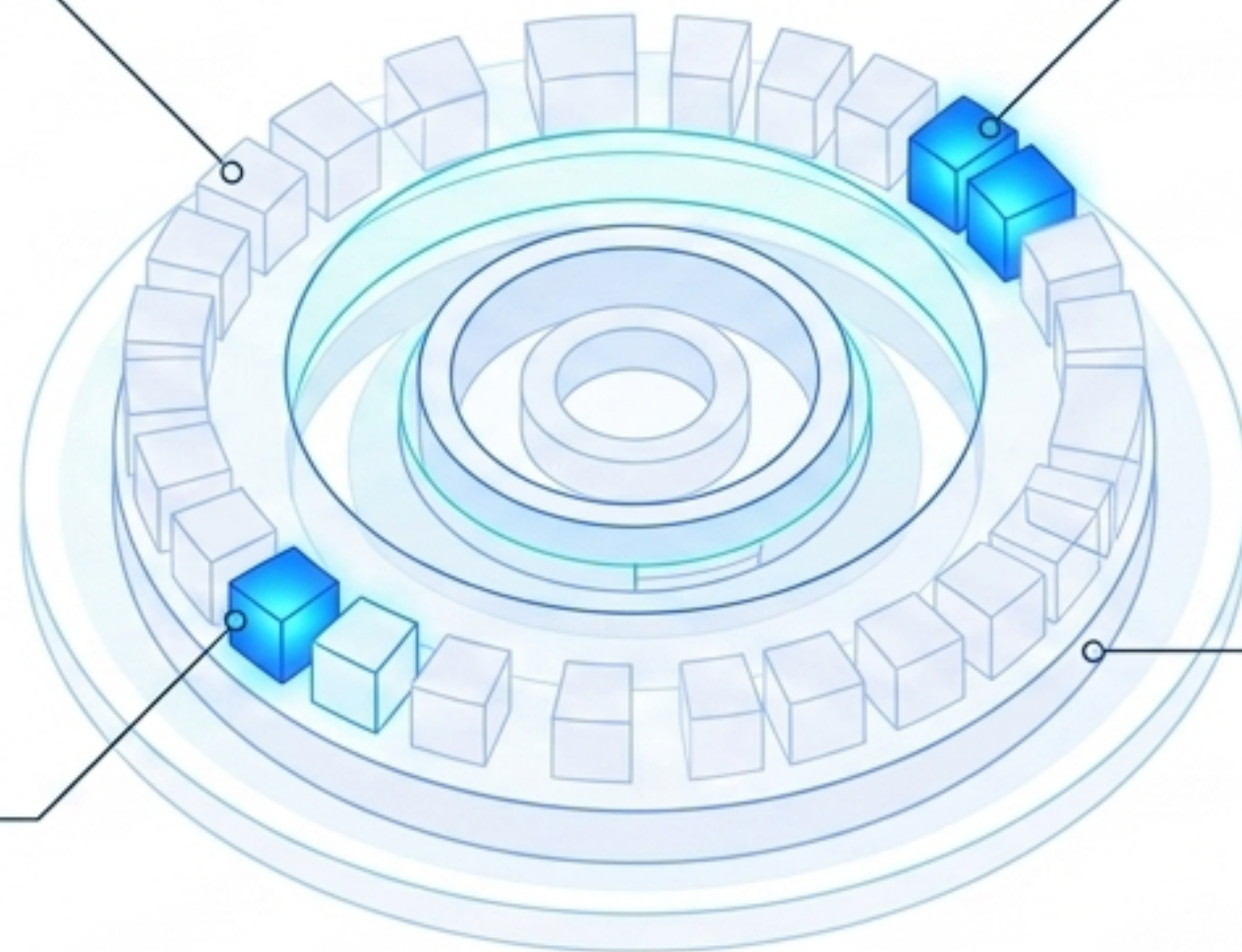
- 専門家 (Experts) の総数：896
- 活性化する専門家：16 (わずか1.8%の稼働率) + 共有専門家2

ハイブリッド・アテンション

- 全93層。ブロックごとに KDA (Kimi Delta Attention): MLA = 3:1 の割合で構成。
- 最大6.3倍の高速デコードとKVキャッシュ最大75%削減を実現。

仕様データ

- 1M (1,048,576) コンテキスト長
- MXFP4 量子化認識学習 (QAT) 適用
公開バリエーション
- 単一 SKU (kimi-k3)。常時推論 (thinking) モデルで、reasoning_effort (low/high/max、既定は max) により調整。旧世代のような Base/Instruct/非 thinking の分割はない¹



世代間比較：K2からの技術的飛躍

	[K2 系 (2025後半)]		[Kimi K3]
パラメータ規模	1兆 (32B活性)	➡	2.8兆 (104B活性)
コンテキスト長	256K	➡	1M (4倍の長文脈処理)
モダリティ	テキスト中心	➡	ネイティブ・マルチモーダル (画像・動画理解対応。MoonViT-V2搭載)
ツール利用	逐次200～300回の呼び出し	➡	常時推論 (Reasoning_effortによる調整)
量子化/最適化	INT4	➡	MXFP4 / SiTU-GLU / Per-Head Muon最適化

コア・インサイト: K3の本質はパラメータの積み増しではない。「KDAとGated MLAの混在」
「LatentMoEによる極疎ルーティング」により、K2比で『約2.5倍のスケーリング効率』を達成した点にある。

ベンチマーク・ダッシュボード：実力の多角的な検証

総合スコア
(Intelligence Index v4.1)

57点 (3~4位)

フロントエンド・コード
(LMArena Frontend Code)

1位 (1,679 Elo)

強み (Strengths)

- オープンウェイト最上位クラス。DeepSeek V4 Pro(44点)やGLM-5.2(51点)を凌駕。
- Fable 5に対するペアワイズ勝率76%。

レッドフラグ (Red Flags)

- コスト高: 1タスクあたり\$0.94 (オープンウェイトとしては高額)。
- 過剰な冗長性: 出力トークン数が中央値の約2倍。
- 処理速度: 1タスク平均 約56分 (Fable 5の約2.5倍の遅さ)。
- 幻覚率の悪化: AA-Omniscienceにおいて精度は向上 (33%→46%) したものの、幻覚率も (39%→51%) へ悪化。

フロンティア・マトリックス：最上位モデルとの競合関係

	K3	Fable 5	GPT-5.6 Sol	Opus 4.8	GPT-5.5
GPQA Diamond	93.5	92.6	94.1	91.0	93.5
HLE-Full (有ツール)	56.0	63.0	58.0	57.9	52.2
AA-LCR (長文脈)	74.7	70.0	73.7	67.7	74.3
MMMU-Pro (無ツール)	81.6	81.2	83.0	78.9	81.2

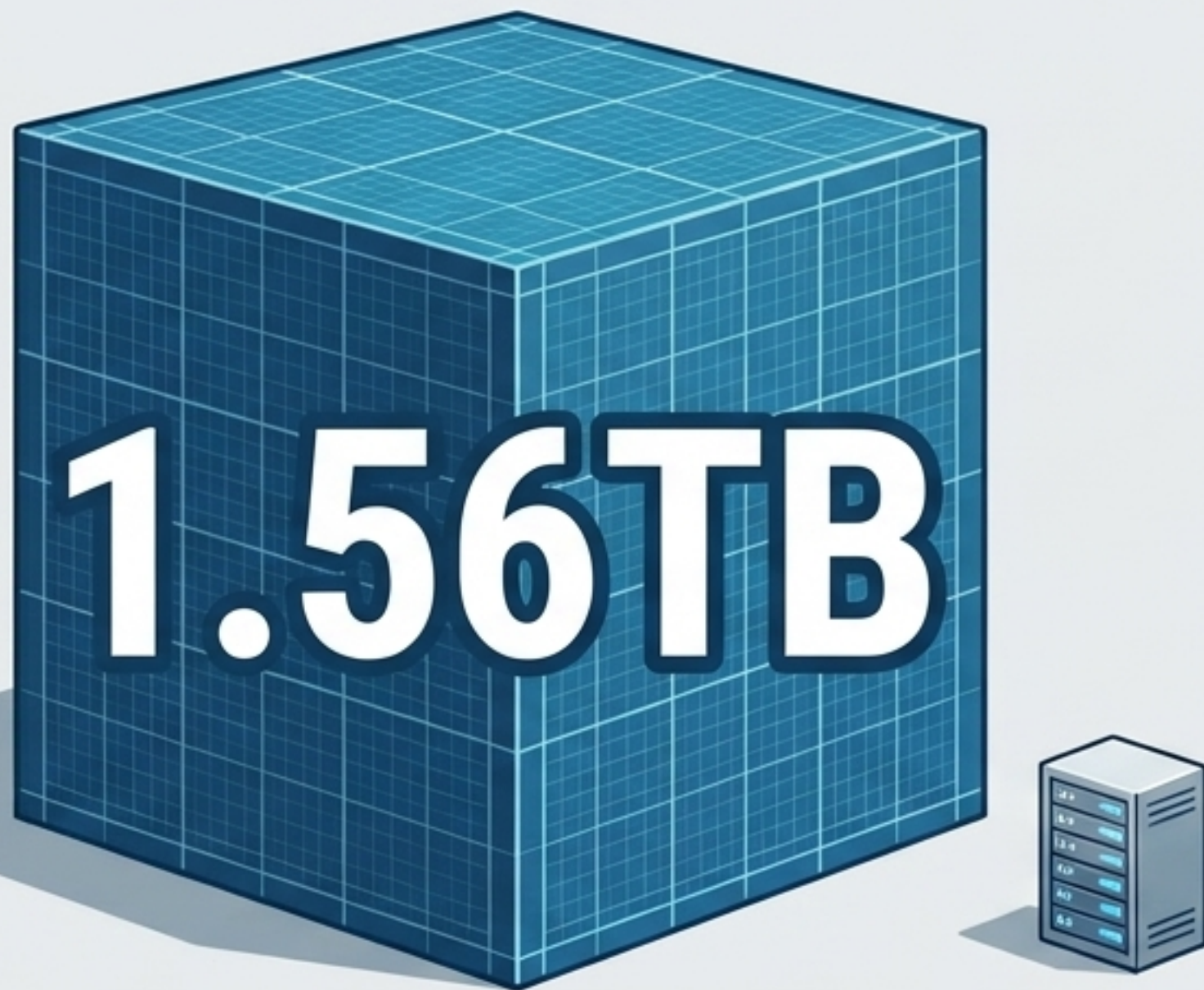
⚠ [CRITICAL WARNING: ハーネス混在の問題]

上表はモデルごとに異なる評価環境（ハーネス）で測定されている。K3は「Kimi Code」、競合は「Claude Code」や「Terminus 2」等を使用。ハーネスの違いだけでスコアが10～26点変動するため、モデル単体の純粋な性能比較としては信頼性が低いとの指摘が相次いでいる。

（※ 「SWE-bench Verified 76.8%」等の出典不明情報にも要注意）

[SUBJECT: INFRASTRUCTURE] | [WEIGHT: 1.56TB]

「オープン」の幻想：実運用ハードウェアの現実



Hugging Face公式リポジトリ容量

要件リスト

最低構成要件: 8 × NVIDIA H100

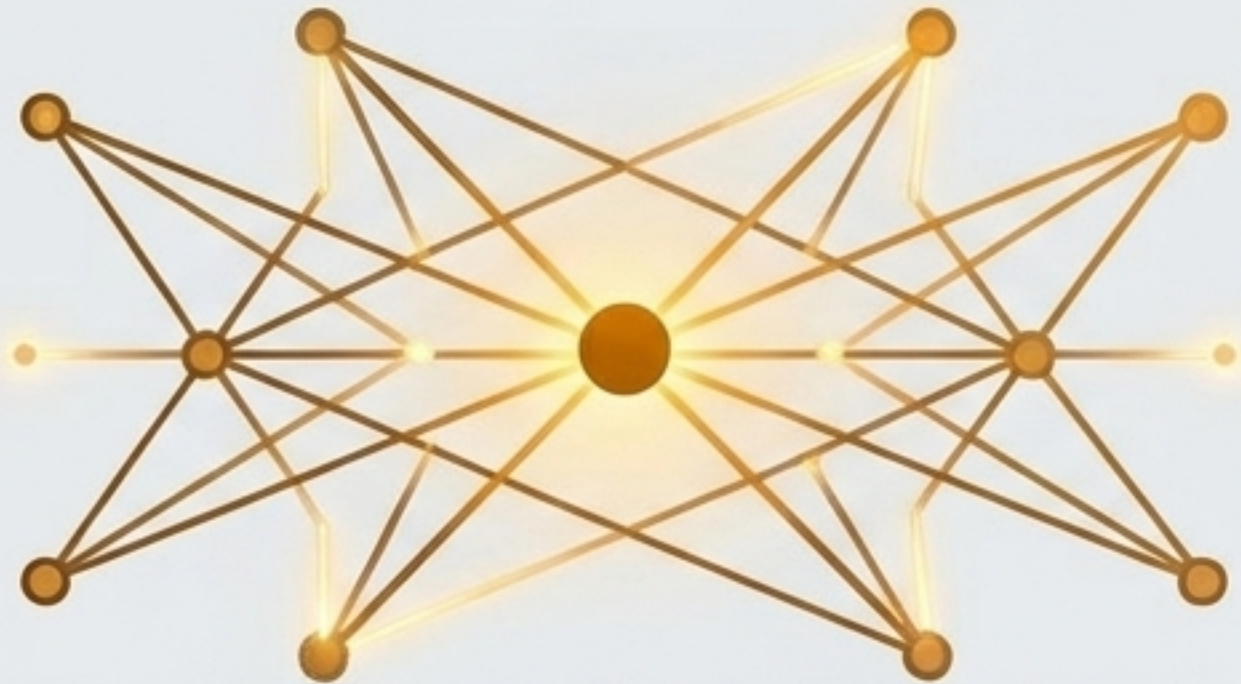
推奨構成要件: 64基以上のアクセラレータ (スーパーノード構成)

量子化 (1-bit GGUF) の限界:
Unsloth等により594~620GBまで圧縮可能だが、それでも650GB超のRAM/VRAM が必要。

結論: オープンウェイトはローカル実行可能を意味しない。今回の公開は、実質的に個人開発者ではなく、資金力のあるクラウド・推論プロバイダ向けである。

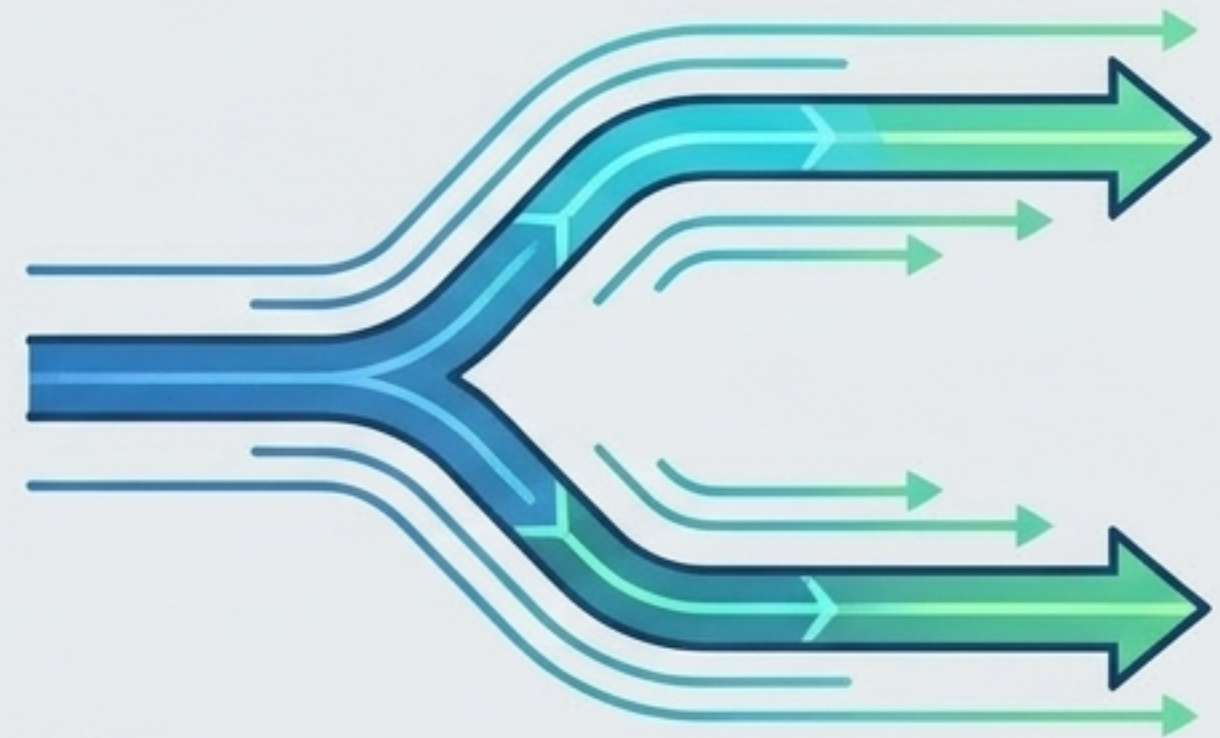
地政学インパクトとエコシステムの分断

⚠ 米中の対立



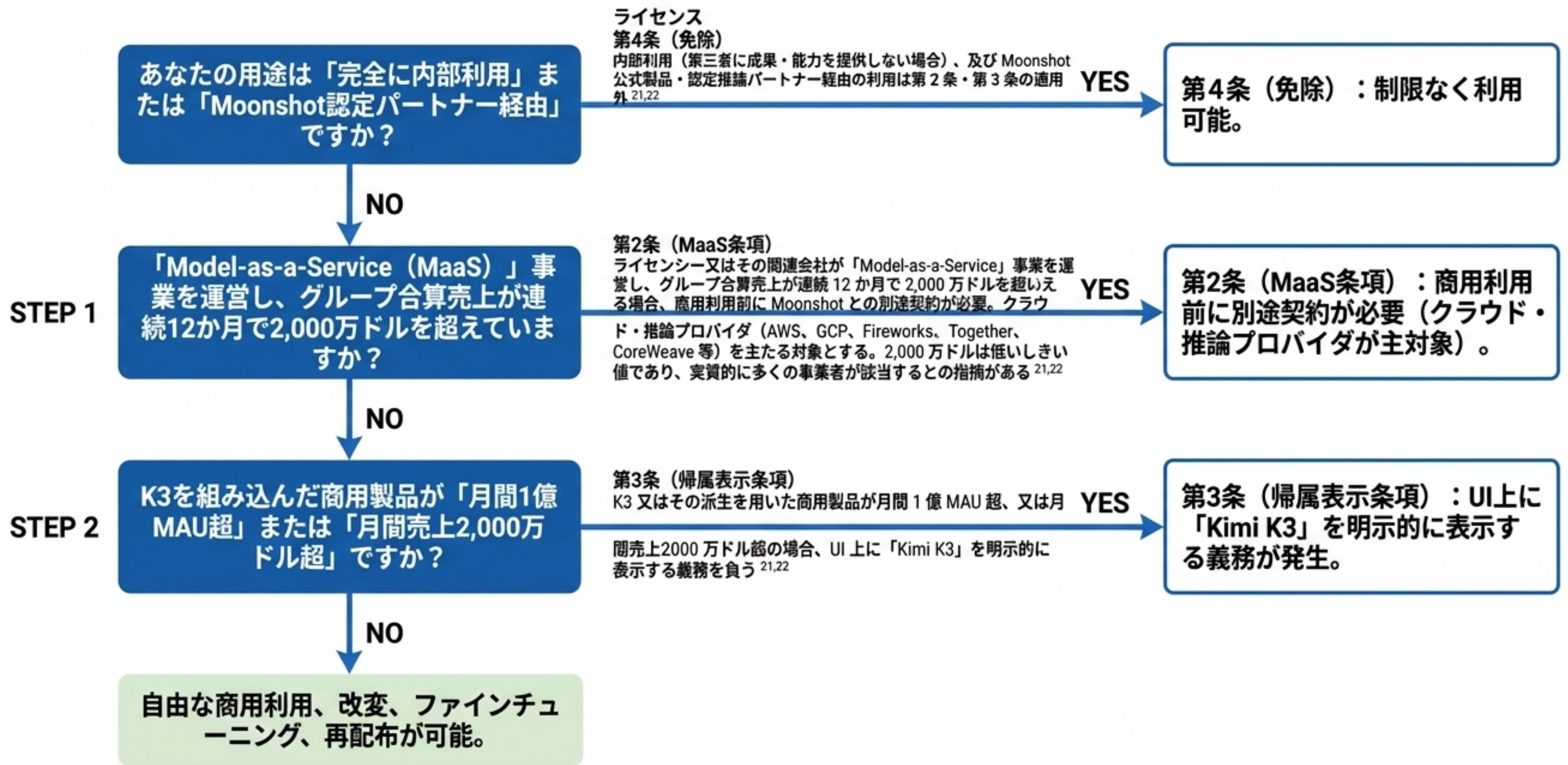
- **蒸留疑惑:** 米ホワイトハウスが、K3はAnthropic「Fable」の蒸留であると主張（証拠未提示）。
- **タイムラインの反証:** Fable公開(7/1)からK3公開(7/16)までわずか15日。「15日でフロンティアモデルを訓練したならギネス級」との専門家の反論。
- **輸出規制:** タイ経由でのNvidia GB300サーバアクセス疑惑。BIS（米商務省産業安全保障局）が調査中。

✅ エコシステムの対応



- **Day-0 ホスティング展開:** Fireworks AI, Together AI, Microsoft Foundry, Huawei (Ascend CANN), Alibaba Cloud.
- **対応凍結・見送り:** AWS Bedrock, Google Vertex（政策リスクを懸念し、大手ハイパースケーラーは当初対応を見送り）。

ライセンス判定ツリー：MITではない独自の「Kimi K3 License」



エンタープライズ・アクションプラン：実務上の推奨戦略

【ADOPT（採用推奨）】

フロントエンドのコード生成、
長文脈（1M）のドキュメント
理解、自動化タスク。

【AVOID（回避推奨）】

罣の多い複雑なデバッグ、単発の
深い推論、幻覚許容度が極めて
低いセキュリティ精査業務。
（※Fable 5やOpus系を優先）

- ☐ APIによる先行評価: 自社ホストは避け、まずは Fireworks / Together / MS Foundry のAPIで検証する。
- ☐ 自社ホストの条件確認: キャッシュヒット率 90%超かつ「厳格なデータ主権要件」がある場合のみ、8xH100以上の自社運用を検討する。
- ☐ A/Bテストの実施: 公開ベンチマーク（ハーネス汚染の可能性あり）を鵜呑みにせず、自社コードベースでのブラインドテストを行う。
- ☐ 地政学リスクの監視: 米BISの調査動向（Entity List入りの可能性）を継続監視し、代替モデルへの切り替えトリガーを事前に定めておく。