

Kimi K3の重み公開に関する調査・分析報告書

エグゼクティブサマリ

- 公開元は中国・北京のMoonshot AI（月之暗面）である。Kimi K3自体は2026年7月16日にAPI・Webサービスとして発表され、モデル重み、GitHubリポジトリ、Hugging Faceモデルカード、47ページの技術報告書が7月27日から28日にかけて公開された。arXivの登録時刻は2026年7月27日16時49分54秒UTC、すなわち日本時間7月28日1時49分54秒であるため、質問文の「7月28日公開」は日本時間基準では妥当である。一方、Moonshot AIの公式Xは「7月27日公開」と表示しており、正確には「7月27日UTC／7月28日JSTの公開」と表現するのが適切である。 ¹
- 公開されたのは「完全なオープンソース」ではなく、主としてモデル重みと推論に必要な構成情報である。2.8兆の総パラメータ、1040億の有効パラメータ、896専門家中16専門家を選択するMoE、100万トークンのコンテキスト、ネイティブ視覚処理、Kimi Delta Attention（KDA）、Attention Residuals（AttnRes）などが開示された。しかし、事前学習データの具体的なデータセット一覧、総トークン数、言語構成、学習FLOPs、完全な学習コード、データの権利処理記録は公開されていない。したがって、推論・改変は可能でも、モデルを同じ条件で一から再学習する意味での再現性は低い。 ²
- 性能は、特に長時間コーディング、ブラウジング、ツール利用、文書・画像処理でフロンティア級だが、最上位の閉鎖モデルを一貫して上回るわけではない。公式表ではFrontierSWE 81.2、SWE-Marathon 42.0、BrowseComp 91.2、MCPMark-Verified 94.5などが強い一方、CritPtやHLE-FullではClaude Fable 5、GPT-5.6 Solに明確に劣る。Moonshot AI自身も「総合性能は最強の独自モデルにまだ及ばない」と認めている。評価ハーンズ、推論努力、ハードウェア、コンテキスト管理方式がモデルごとに異なるため、表の数値を完全な同条件比較とみなすべきではない。 ³
- 公開の実用価値は大きい、セルフホストの敷居も極めて高い。重みは約1.5TB級で、公式は64基以上のアクセラレーターを持つスーパーノードを推奨している。つまり「誰でもダウンロードできる」と「一般企業や個人が安価に運用できる」ことは同義ではない。公開直後5日程度で、Hugging Face上では約56万ダウンロード、約9360件の「いいね」、28の派生ファインチューニング、27の量子化、16のSpacesが確認され、公式Xの公開投稿は約1270万表示に達した。 ⁴
- 最大の留保点は、ライセンス、データ透明性、安全性文書の不足である。小規模・中規模の一般的な商用利用は広く認められる一方、一定規模を超えるModel-as-a-Service事業者にはMoonshot AIとの別契約が必要であり、大規模サービスには「Kimi K3」の表示義務がある。技術報告書には独立した「Safety」「Limitations」節や人口統計的バイアス評価が見当たらず、UK AISI／米国CAISIは、K3が最先端の米国モデルより低いもののGLM-5.2より高いサイバー能力を持ち、標準の安全策が攻撃支援の試行を阻止しなかったと報告している。 ⁵

前提・公開日の確定と公式資料

調査範囲と不明点

本報告の情報基準日は2026年8月1日、日本時間である。一次資料としてMoonshot AI公式ブログ、公式X、GitHub、Hugging Faceモデルカードとライセンス、arXiv技術報告書を優先し、次に政府研究機関、独立評価機関、主要メディア、技術者ブログ、SNS・コミュニティ投稿を参照した。

「Kimi K3の重み」という表現は、Hugging Face等で配布される学習済みパラメータ群を指すものと解釈した。公開ファイルの**最初のアップロード時刻を秒単位で確定できる完全なHugging Faceコミット履歴**は今回取得できなかったため、公開日は、公式X、arXiv登録時刻、日本語メディアの確認時刻を組み合わせで判定している。

現時点で不明または未公開なのは、主に次の事項である。

項目	公開状況	評価
モデル重み	公開	Hugging Faceから取得可能
推論構成・アーキテクチャ	おおむね公開	実装・推論の再現は可能
技術報告書	公開	arXiv v1、47ページ
事前学習データセットの個別名称	未公開	出所・権利・偏りの監査が困難
総学習トークン数	未公開	規模・効率の第三者検証が困難
学習FLOPs・総計算量	未公開	EU AI Act上のシステミックリスク判定にも影響
学習用クラスタの完全仕様	未公開	システムレベルの再現性が不足
完全な事前学習コード	未公開	重み公開と同一モデル再学習は別問題
日本語専用ベンチマーク	未公開	日本語実務性能は自社評価が必要
人口統計的・文化的バイアス評価	未公開	公平性評価不能
包括的なモデル安全性報告書	未公開	攻撃性、欺瞞、CBRN等の横断評価は確認不能

技術報告書は、学習データを「Web Text、Code、Mathematics、Knowledge」と大規模視覚コーパスに分類し、ルール、分類器、重複除去によるフィルタリングを説明している。しかし、データセット名、ライセンス構成、著作権者別内訳、言語別比率、個人情報除去率は示していない。 ⁶

公開タイムライン

```
timeline
    title Kimi K3発表から重み公開まで
    2026-07-16 : Moonshot AIがKimi K3を正式発表
                : Kimi.com・Kimi Code・APIで提供開始
                : 重みを7月27日までに公開すると予告
    2026-07-22 : vLLM・NVIDIA・AMD等がDay-0対応を準備
    2026-07-23 : UK AISI・米国CAISIが予備サイバー評価を公表
    2026-07-27 UTC : 公式Xが重みと技術報告書の公開を告知
                : GitHub・Hugging Faceで配布
    2026-07-27 16:49 UTC : arXiv技術報告書v1登録
    2026-07-28 01:49 JST : 日本時間では7月28日の公開
                : 日本語・英語メディアの公開報道が集中
    2026-07-28以降 : 推論実装・量子化・アーキテクチャ分析が拡大
```

7月16日の公式ブログは、重みを「2026年7月27日までに公開する」と予告していた。公式Xの重み公開投稿は7月27日付であり、arXivは7月27日16時49分54秒UTCに登録された。これは日本時間では7月28日1時49分54秒となる。 ⁷

したがって、今後の引用や社内記録では、単に「7月28日公開」とするより、「2026年7月27日UTC／7月28日JSTに重み・技術報告書を公開」と記載するのが最も厳密である。

公式原文と和訳

一次資料	重要な原文抜粋	和訳・意味
Hugging Faceモデルカード	“Kimi K3 is an open-weight, native multimodal agentic model.”	「Kimi K3は、オープンウェイトで、ネイティブなマルチモーダル・エージェントモデルである。」重み公開、視覚入力、ツール利用を中核に据える定義である。 ⁸
arXiv技術報告書	“We release the full Kimi K3 model weights...”	「将来の研究と普及を促進するため、Kimi K3の完全なモデル重みを公開する。」ここでいう“full”は重み全体であり、学習データや完全な訓練環境まで含むとは書かれていない。 ⁹
公式ブログ	“its overall performance still trails the most powerful proprietary models”	「総合性能は、最も強力な独自モデルにまだ及ばない。」公式自身がClaude Fable 5とGPT-5.6 Solへの劣後を認めており、発表は単純な“世界最高”主張ではない。 ¹⁰
技術報告書・学習データ	“Web Text, Code, Mathematics, and Knowledge”	「ウェブテキスト、コード、数学、知識」。これに視覚コーパスを加えるが、具体的なデータセット一覧はない。 ⁶
ライセンス	“must enter into a separate agreement with Moonshot AI”	「Moonshot AIとの別個の契約を締結しなければならない。」一定売上を超えるMaaS事業者に適用される重要な商用条件である。 ¹¹
NIST／CAISI	“safeguards did not prevent it from attempting cyber exploit development”	「安全策は、サイバー攻撃コードの開発を試みることを阻止しなかった。」能力の高さだけでなく、公開時の安全制御の限界を示す。 ¹²

技術・性能・再現性の分析

アーキテクチャ

Kimi K3は、総パラメータ2.8兆、各トークンで有効になるパラメータ約1040億のMixture-of-Expertsモデルである。有効比率は単純計算で総数の約3.7%であり、896のルーティング専門家から16を選択し、さらに2つの共有専門家を利用する。層数は93で、69層がKDA、24層がGated MLA、隠れ次元7168、語彙16万、視覚エンコーダーMoonViT-V2は約4.01億パラメータである。¹³

構成項目	Kimi K3	技術的意味
総パラメータ	2.8兆	知識・専門化容量は非常に大きい
有効パラメータ	1040億／トークン	密モデル2.8兆より計算量を削減
専門家数	896	高い専門化が可能だが、ルーティングと通信が複雑
選択専門家	16＋共有2	高スパース性と一定の共通知識を両立
層数	93	69 KDA＋24 Gated MLA

構成項目	Kimi K3	技術的意味
コンテキスト	1,048,576トークン	長大なコードベース、文書、映像系列を想定
視覚エンコーダー	MoonViT-V2、401M	画像・動画をテキストと統合
重み・活性値	MXFP4／MXFP8	量子化を後付けではなく学習時から考慮
推奨構成	64アクセラレーター以上	個人向けローカルモデルではない

数値は公式モデルカードと技術ブログによる。¹⁴

KDAは、フルアテンションだけに依存せず、リカレントな状態更新を用いるハイブリッド線形アテンションである。AttnResは、全層の情報を一様に残差加算するのではなく、深さ方向の過去表現を選択的に取り出す。Stable LatentMoE、Quantile Balancing、Per-Head Muonなどは、896専門家を持つ極端に疎なモデルの学習安定性と負荷均衡を支える設計である。¹⁵

Sebastian Raschkaは、K3をKimi Linearの「**scaled-up production version**」、すなわち「大規模化された実用版」と整理している。これは完全に未知の方式というより、48B級で実証されたKDAを2.8T級へ拡張し、MoE、視覚、エージェントRLを統合したものと理解するのが適切である。¹⁶

なお、Kimi Linearの先行研究では、特定条件下で最大75%のKVキャッシュ削減、100万トークン生成時に最大約6倍のデコード高速化が報告されているが、これは48B級の先行モデルに関する結果である。**同じ倍率がK3の2.8T実運用で得られると推定してはならない。**¹⁷

ネイティブマルチモーダルと長文脈

K3は、言語モデルの学習後に視覚エンコーダーを接続する方式ではなく、学習開始時から視覚トークンとテキストトークンを共同最適化する「ネイティブマルチモーダル」方式を採用。画像、動画、OCR、コードとレンダリング結果を同じトークン系列内で扱うため、コードを書き、スクリーンショットを確認し、結果を修正する反復型エージェントに向く。¹⁸

100万トークン対応は、最初から全学習を100万トークンで実施したという意味ではない。事前学習では8Kから64Kへ、クールダウン段階では256Kから1Mへ段階的に拡張している。明示的な位置埋め込みを使用しないNoPEとKDAのゲート・減衰構造を用い、長文書や動画だけでなく、情報を長距離に分散した合成タスクも追加している。¹⁹

これは計算効率上合理的だが、「100万トークン入力可能」と「100万トークン全域で均一に正確な検索・推論が可能」は別である。公式表のBrowseCompでは、30万トークン時点でコンテキスト圧縮を起動する構成が91.2で、完全な100万トークンを圧縮なしで使う場合は90.4だったと説明されている。運用側のコンテキスト管理手法が結果に影響する。²⁰

ベンチマーク比較

以下は公式モデルカードから代表的な評価を抜粋したものである。太字は表中の最高値であり、異なる評価ハーネスを横断した公式集計である点に注意が必要である。

ベンチマーク	Kimi K3	Claude Fable 5	GPT-5.6 Sol	Claude Opus 4.8	GPT-5.5	GLM-5.2
GPQA Diamond	93.5	92.6	94.1	91.0	93.5	91.2

ベンチマーク	Kimi K3	Claude Fable 5	GPT-5.6 Sol	Claude Opus 4.8	GPT-5.5	GLM-5.2
CritPt	23.4	28.6	32.3	20.9	27.1	20.9
AA-LCR	74.7	70.0	73.7	67.7	74.3	71.3
DeepSWE	67.5	70.0	73.0	59.0	67.0	46.2
FrontierSWE	81.2	86.6	71.3	66.7	64.9	67.3
SWE-Marathon	42.0	35.0	39.0	40.0	14.0	13.0
BrowseComp	91.2	88.0	90.4	84.3	84.4	未掲載
MCPMark-Verified	94.5	87.4	92.9	76.4	92.9	未掲載
AutomationBench	30.8	29.1	29.7	27.2	22.7	12.9
AA-Briefcase Elo	1548	1583	1495	1354	1158	1260
OSWorld-Verified	84.8	85.0	83.0	83.4	79.0	未掲載
OmniDocBench	91.1	89.8	85.8	87.9	89.4	未掲載
MMMU-Pro	81.6 / 83.4	81.2 / 86.5	83.0 / 84.6	78.9 / 82.7	81.2 / 83.2	未掲載

公式モデルカードによる。²¹

結果からは、次の傾向が読み取れる。

強い領域は、長期間のソフトウェア作業、ブラウジング、MCP・ツール操作、業務自動化、文書視覚処理である。SWE-Marathon、BrowseComp、MCPMark、AutomationBench、OmniDocBenchでは比較対象中の最高値を示す。単発の知識問題だけでなく、複数ステップを持続し、外部ツールを操作する設計の効果と整合する。²²

一方、**研究水準の難問推論では一貫した優位性はない。**CritPtはGPT-5.6 Solの32.3に対して23.4、HLE-FullもClaude Fable 5やGPT-5.6 Solを下回る。DeepSWEでもGPT-5.6 SolとClaude Fable 5に及ばない。公式ブログと論文が「総合では最上位の独自モデルに及ばない」とする評価は、個別スコアともおおむね一致する。²³

ただし、評価は完全な同条件ではない。K3にはKimi Codeハーネス、GPT系にはCodexハーネス、Claude系にはClaude Codeハーネスが使われた項目があり、推論努力も“max”“xhigh”“fallbackあり”など異なる。一部のK3評価はH100ではなくH20で3回平均され、別の値はArtificial Analysisや公式リーダーボードから引用されている。したがって、差が1〜3ポイント程度の項目ではモデル自体の差とハーネス差を切り分けられない。²⁰

公開前の独立評価では、Artificial AnalysisがK3にIntelligence Index 57、AA-Briefcase Elo 1543を付与し、K2.6から727 Elo改善したと報告した。公式モデルカードの後日の値1548とはスナップショット時点が異なる。これは独立した強い裏付けではあるが、重み公開後に第三者が同一ハードウェアで再実行した完全な検証ではない。²⁴

学習データ、バイアス、安全性

技術報告書が開示する事前学習データは、ウェブテキスト、コード、数学、知識、画像キャプション、画像・テキスト混在文書、OCR、知覚、動画、視覚コーディングなどの大分類にとどまる。オープンソースの視覚データと社内合成パイプラインを組み合わせ、ルール・分類器・重複除去を適用したことは説明されるが、収集元、各データセットの比率、著作権状態、拒否・削除手続は不明である。⁶

そのため、現時点では以下を検証できない。

- ・日本語、中国語、英語その他の言語別性能・偏り
- ・性別、人種、宗教、国籍、政治的観点に関するステレオタイプ
- ・医療・法務・金融領域の誤情報率
- ・学習データ中の個人情報・機密情報の混入率
- ・著作権者による機械学習利用拒否の反映状況
- ・特定閉鎖モデルの出力が学習にどの程度含まれるか

技術報告書の全文検索では、独立した“Safety”または“Limitations”節は確認できなかった。“bias”という語は主にMoEルータの数理的なバイアス値を指し、社会的バイアスの評価ではない。これは「安全性が評価されていない」と断定する証拠ではないが、**公開された主要文書だけでは包括的安全評価を監査できないこと**を意味する。²⁵

UK AISIと米国CAISIの予備評価では、K3はExploitBenchで32%、GLM-5.2は24%だった。K3は任意コード実行に至った例が0/41で、最も高性能な比較モデル群の平均20/41より低かったが、32段階の模擬企業ネットワーク攻撃では平均17段階まで進み、10回中1回は全工程を完了した。GLM-5.2の平均は11段階、最先端米国モデル群は28.5段階だった。²⁶

この評価は「K3が最も危険」という結果ではない。むしろ、**閉鎖型最先端モデルより能力は低い**が、**公開重みモデルとしては明確に能力水準が上昇し、既定の安全策も攻撃支援を完全には止めない**という二面的な結果である。評価は少数のベンチマークに限られ、比較対象の閉鎖モデルはシステム安全策を無効化した条件で測定されているため、公開サービス同士の安全性比較として直接解釈すべきではない。¹²

再現性の水準

K3の公開は、次の三段階で区別すべきである。

再現対象	水準	理由
同じ重みで推論する	高い	重み、設定、トークナイザー、推論支援が公開
推論エンジンを独自実装する	中～高	アーキテクチャの記述は詳細だが、KDA等の最適化が必要
公式ベンチマークを同条件で再実行する	中程度	ハーネスや外部環境、推論予算の差がある
同じモデルを一から再学習する	低い	データ、総トークン、計算量、完全コードが未公開
データの法的・倫理的監査	低い	出所・ライセンス・個人情報処理が不明

Hacker Newsでは「**The architecture is reproducible, not the model weights**」という表現をめぐり、重みの再利用、アーキテクチャの再実装、訓練の再現が混同されているとの議論が起きた。厳密には、重みは

複製可能であり、アーキテクチャも実装可能だが、**学習過程そのものは再現不能**という整理が最も正確である。²⁷

メディア・コミュニティの反響と定量指標

時系列で見た主要反応

日付	出来事・反響	主な論点
7月16日	Moonshot AIがK3を発表。 TechCrunchなどが速報	2.8T、1Mコンテキスト、重み公開予告、低価格APIが注目された。 ²⁸
7月17日	Reutersが世界最大級のオープンウェイトモデルとして報道	米中のモデル性能差が縮小したとの地政学的解釈が広がった。 ²⁹
7月18日	Axios等が「AI競争を変えるモデル」と評価	技術評価と米中競争の物語が結び付けられた。 ³⁰
7月20日	需要急増で新規サブスクリプションを一時停止と報道	モデル人気と同時に、計算資源不足・供給制約が可視化された。 ³¹
7月20日以降	オープンウェイト政策をめぐる英語圏の議論が拡大	中国企業が重み公開を国際普及戦略に使っているとの見方が強まった。 ³²
7月22日	vLLMが本番規模サポートのプレビューを公開	vLLM、Moonshot AI、NVIDIA、AMDが公開初日からの推論対応を準備。 ³³
7月23日	UK AISI／米国CAISIがサイバー能力評価を公開	GLM-5.2以上、最先端米国モデル以下、安全策の不十分さが焦点となった。 ²⁶
7月23～24日	米政府関係者が閉鎖モデルからの不正蒸留を主張	Moonshot AIは否定。公開された決定的証拠はなく、技術者から反論も出た。 ³⁴
7月27日	公式Xが重み・技術報告書の公開を告知	公式投稿は取得時点で約1270万表示。PC Watchは日本で公開予定を報道。 ³⁵
7月28日	arXiv、日本語・英語メディアの記事が集中	ExaWizards、ASCII、Tom's Hardware、The Verge等が規模、性能、公開戦略を解説。 ³⁶
7月28日	Raschkaがアーキテクチャ詳細を分析	Kimi Linearからの拡張、LatentMoE、推論効率が評価された。 ¹⁶
7月30日以降	Zenn、Gizmodo Japanなどで日本語解説が増加	「オープンウェイトとオープンソースの違い」「一般PCで動かせない規模」が主要論点となった。 ³⁷
8月1日	一般向け日本語報道にも波及	技術者向けニュースから、米中AI競争・産業影響を扱う一般報道へ拡大した。 ³⁸

メディアの評価軸

英語主要メディアでは、大きく三つの物語が形成された。

第一は、**閉鎖型フロンティアモデルとの性能・価格競争**である。Reuters、TechCrunch、Tom's Hardwareなどは、K3がClaude Fable 5やGPT-5.6 Solに近い性能を、より低いAPI価格または自己運用可能な形で提供する

点を重視した。Tom’s Hardwareは重み公開後、閉鎖モデルに対する「shot across the bow」と表現した一方、学習データと方法が依然として非公開であることも指摘した。 39

第二は、**中国のオープンウェイト戦略**である。The Vergeは、中国企業による重み公開が単なる研究公開ではなく、世界の推論基盤、ツール、開発者エコシステムへ中国モデルを浸透させる産業戦略として機能すると分析した。Reuters Breakingviewsも、オープンウェイトモデルが閉鎖モデルに追い付きつつある一方、ハードウェア、クラウド、輸出管理への依存は残ると論じた。 40

第三は、**米中対立、蒸留、輸出管理**である。米政府関係者はK3開発にAnthropicモデルの不正な大規模蒸留が使われたと主張し、制裁の可能性にも言及した。Moonshot AIはこれを否定し、研究者の一部は公開証拠の不足や開発時系列の不自然さを指摘した。2026年8月1日時点で、司法判断、規制当局の確定認定、公開された技術的証拠は確認できず、**事実としてではなく係争中の疑惑として扱う必要がある**。 41

日本語メディア・ブログでは、地政学よりも「2.8兆のうち1040億だけが稼働する仕組み」「約1.5TBの重み」「64基以上のアクセラレーター」「オープンウェイトとオープンソースの違い」が中心だった。ExaWizardsはMoEを「案件ごとに適任者を選ぶ巨大企業」に例えて解説し、Zennの記事は学習データや完全な訓練過程が開示されていない点を明示した。 42

SNS・技術コミュニティ

公式Xの重み公開投稿は、取得時点で**約1270万表示**となっており、公開直後の関心規模を示す最も大きな単一指標である。ただし、表示数は閲覧者数や実利用者数と同義ではなく、重複閲覧やニュース拡散も含む。 43

Redditのr/LocalLLaMAでは、発表直後から「どの量子化なら保存できるか」「A100、H200、B300で何台必要か」「約1.4TBをどう配布・ロードするか」という運用面の議論が中心となった。ある短い反応の“**Gonna need an upgrade**”は、公開の自由度と一般ユーザーのハードウェア能力との隔たりを端的に示している。別スレッドでは、A100上でのメモリ要件を事前計算する試みも投稿されたが、公開前の推測値を含むため公式仕様と区別すべきである。 44

Hacker Newsでは、KDA、AttnRes、LatentMoEの設計そのものに加え、「重み公開を再現性と呼べるか」「アーキテクチャ仕様だけで研究再現と言えるか」が論点になった。議論は概して、モデルを動かせることを高く評価しつつ、訓練データと計算過程がないため科学的再現性は限定的、という方向に収束している。 27

日本語圏では、Zenn、note、企業技術メディアの記事は確認できたが、本調査の検索範囲では、Qiitaや日本の匿名掲示板において、一次資料や実測結果に基づく高シグナルの投稿はまだ限定的だった。これは投稿が存在しないという意味ではなく、公開から約5日しか経過しておらず、実機検証に必要なダウンロード時間とインフラ規模が大きいことも影響していると考えられる。

定量的な反響指標

以下は2026年8月1日前後に取得した動的スナップショットである。プラットフォームごとに計測定義が異なるため、相互に単純比較できない。

指標	観測値	解釈・留保
公式X・重み公開投稿の表示	約1270万	世界的認知を示すが、利用者数ではない。 43

指標	観測値	解釈・留保
Hugging Faceダウンロード	559,924／直近1か月表示	公開後数日なので急速な取得を示すが、分割ファイル取得や自動処理の影響を含み得る。 ⁴⁵
Hugging Face「いいね」	約9,360	関心・ブックマークに近い指標。 ⁴⁶
Hugging Face派生ファインチューニング	28	公開直後から派生開発が開始。 ⁴⁵
Hugging Face量子化	27	ハードウェア要件を下げる試みが活発。 ⁴⁵
Hugging Face Spaces	16	デモ・推論統合の初期形成。 ⁴⁵
GitHub フォーク	約540	コード・文書の派生作業を示す。取得時点により542～544と変動。 ⁴⁷
GitHub Issues／PR	9／7	公開直後の統合・不具合修正段階。 ⁴⁸
重みデータ量	約1.5～1.56TB	取得・保管・ロード自体が大規模運用課題。日本語記事による集計値。 ⁴⁹
本調査の採択反響ソース	34件	公式、主要メディア、技術ブログ、高シグナル投稿の限定サンプルであり、全ウェブ記事数ではない

Hugging FaceやXから過去の日別値を復元できる公式公開API履歴は本調査では得られなかった。このため、以下のグラフはダウンロード数や表示数の推移ではなく、**本調査で採択した日付確認可能な公式発表・記事・技術ブログ・コミュニティ投稿34件の日別公開数と累積数**を示す。7月16日の製品発表と7月27～28日の重み公開に二つの報道ピークがある。

Kimi K3に関する選定ソースの日別公開数と累積数

このグラフは報道量の絶対値ではない。検索エンジンの索引、言語、媒体選定による標本偏りがあり、SNS投稿の全数、転載記事、動画、非公開コミュニティは含まない。一方、製品発表時と重み公開時に関心が二段階で立ち上がったという時系列構造は、公式発表、Reuters、vLLM、PC Watch、Tom's Hardware、日本語技術記事の公開日と整合する。 ⁵⁰

法的・倫理的評価

Kimi K3ライセンス

Kimi K3 Licenseは、モデル重み、パラメータ、構成ファイル、推論・訓練コード、文書をまとめて“Software”と定義し、使用、複製、改変、統合、公開、配布、サブライセンス、販売、実行、デプロイ、ファインチューニング、派生物作成を広く認めている。ただし、以下の独自条件がある。 ¹¹

利用形態	条件・可否
研究、評価、非商用利用	原則可能。著作権表示・許諾表示を維持し、法令を遵守する
社内利用	原則可能。第三者にモデル、出力、基盤能力を提供しない社内利用は特定の大規模事業条件から除外

利用形態	条件・可否
通常の商用製品への組み込み	原則可能。ただし大規模サービスの表示条件に注意
ファインチューニング・派生モデル	原則可能
再配布・販売	原則可能。ただしライセンス表示等を継承
MaaS・推論API	売上規模によって別契約が必要
大規模消費者サービス	「Kimi K3」の目立つ表示が必要となる場合がある
保証	なし。「AS IS」で提供され、損害責任も原則否認

最も重要なのは第2条である。ライセンサーまたは関連会社がModel-as-a-Service事業を運営し、ライセンサーと関連会社の合算売上が**連続する12か月で2000万米ドルを超える場合**、K3または派生物を商用目的で使う前にMoonshot AIとの別契約が必要になる。ここでのMaaSは、第三者が入力、パラメータ、訓練データを実質的に制御できる形で、推論またはファインチューニング機能を提供する事業を指す。¹¹

第3条では、K3または派生物を利用する商用製品・サービスが、**月間アクティブユーザー1億人超、または月間売上2000万米ドル超**の場合、ユーザーインターフェース上に“Kimi K3”を目立つ形で表示することを求める。社内利用、Moonshot AI公式製品、認定推論パートナー経由の利用は、第2条・第3条の対象外とされる。¹¹

したがって、「商用利用可能」は概ね正しいが、「あらゆる企業が無条件・無制限に商用利用可能」は誤りである。特にクラウド推論、ホスティング、ファインチューニング代行、基盤モデルAPIを提供する企業は、売上判定に関連会社を含む点、閾値を超えた後ではなく**商用利用前に別契約が必要と読める点**を法務確認すべきである。

オープンウェイトとオープンソース

Open Source Initiativeは、モデルのパラメータにアクセスできる状態を「open weights」と区別し、完全なオープンソースAIには、モデルの研究、使用、改変、共有を可能にする十分なコード、データ情報、権利が必要だとしている。⁵¹

K3は重み、構成、技術文書を公開しており、一般的なモデル公開としてはかなり開放的である。しかし、訓練データと完全な訓練コードがなく、売上規模・事業形態による追加契約条件もある。OSIのOpen Source Definitionは利用分野による差別や特定事業への制限を認めないため、**Kimi K3 LicenseがOSI準拠のオープンソースライセンスと認定される可能性は低い**と考えられる。これは本報告のライセンス文言とOSI原則に基づく分析であり、OSIによるK3個別の正式認定結果ではない。⁵²

実務上は、調達文書や広報で「オープンソースKimi K3」と一括表示するより、「**独自ライセンスで公開されたオープンウェイトモデル**」と記載する方が正確である。

学習データと知的財産

公式報告書は、オープンソースの視覚コレクション、ウェブテキスト、コード、数学、知識データ、社内合成データを使ったと説明するが、権利者別・ライセンス別の詳細を公開していない。このため、利用企業が重みだけから学習データの適法性を独立監査することはできない。⁶

米政府関係者は、Moonshot AIがAnthropicのFableモデルを大規模に蒸留してK3を開発したと主張し、制裁の可能性にも言及した。Moonshot AIは否定し、複数の研究者は、公開証拠がないことや開発時系列から単純な「Fable 5の複製」とする説明に疑問を呈した。2026年8月1日時点では、疑惑の対象となったプロンプト数、出力数、使用時期、SFT・RL・評価のどの段階に使われたかは公表されていない。 53

この問題は、重みをダウンロードする一般利用者が直ちに侵害責任を負うと決まったことを意味しない。ただし、将来、制裁、Entity List指定、クラウド配布制限、知的財産訴訟が現実化した場合、モデルの更新、サポート、決済、米国系GPU・クラウドでのホスティングに影響する可能性がある。企業採用では「技術性能」だけでなく、供給継続性と地政学的コンプライアンスを別枠で評価すべきである。 54

EU AI Act

EUでは、汎用AIモデル提供者の義務が2025年8月2日から適用されており、2026年8月2日から欧州委員会の完全な執行権限、罰金を含む運用が始まる。K3の重み公開は2026年7月27～28日なので、EU市場に提供される場合、原則として新しいGPAI規制環境下で扱われる。 55

一般目的AIモデル提供者には、技術文書、下流提供者向け情報、EU著作権法に従う方針、学習コンテンツの十分に詳細な公開要約、域外事業者の場合のEU代理人などが求められる。自由かつオープンソースのモデルには一部免除があるが、著作権方針と学習内容要約は原則として残り、システミックリスクモデルにはオープンソース免除が適用されない。 56

K3の独自ライセンスが、EUガイドラインのいう「真に自由かつオープンソース」の条件を満たすかは不明である。収益規模に応じた別契約義務があるため、免除の適用を当然視すべきではない。また、システミックリスク推定に使われる学習計算量が公開されていないため、パラメータ数だけから法的分類を確定できない。Moonshot AIまたはEUでK3を自己ブランド提供する事業者は、EU AI Officeとの個別確認が必要となる。 56

日本での利用

日本では、総務省・経済産業省の「AI事業者ガイドライン」が、AI開発者、提供者、利用者に対し、人間中心、安全性、公平性、プライバシー、セキュリティ、透明性、アカウントビリティ等のガバナンスを求めている。2026年3月には改訂版が公表されている。K3をセルフホストしても、導入企業自身の安全管理、ログ、説明、権限管理、事故対応責任が消えるわけではない。 57

個人情報保護委員会は、生成AIサービスへ個人データを入力する場合、利用目的の範囲内であること、提供先の取扱い、学習への利用、外国移転等を確認する必要があると注意喚起している。セルフホスト型K3は外部APIへの送信を減らせる可能性がある一方、モデルへの入力、ログ、ベクトルデータベース、ファインチューニングデータを運用企業が自ら管理する責任が増える。 58

日本の著作権法上、情報解析のための著作物利用に一定の柔軟性があるとしても、生成出力による既存作品との類似、営業秘密、個人情報、契約上の利用制限、国外サービスへの提供は別問題である。特にコード生成、文書生成、画像生成を顧客向け製品へ組み込む場合、出力フィルタ、類似性検査、権利者申告窓口を用意することが望ましい。

倫理的リスク

K3の倫理的評価では、次の非対称性が重要である。

公開により改善する点	公開だけでは解決しない点
重みを検査・改変できる	学習データの出所・同意は分からない
独立安全評価を実施できる	初期の安全策を除去することも可能
閉鎖APIへのデータ送信を避けられる	自社運用側に個人情報・セキュリティ責任が移る
モデル挙動を監査・再現できる	社会的バイアスの原因データは監査できない
研究者が防御策を開発できる	サイバー攻撃、詐欺、監視への転用も容易になる
ベンダーロックインを減らせる	64基以上のアクセラレーターにより計算資源集中は残る

すなわち、オープンウェイトは透明性を高めるが、安全性を自動的に高めるものではない。K3の場合、特にエージェント能力、長時間タスク、端末操作、ブラウジング、コード生成が強いため、単純なチャットボットよりも、ツール権限、ネットワーク権限、認証情報、実行環境の隔離が重要になる。UK AISI/CAISIの結果は、この運用上の注意を裏付ける。²⁶

結論と実務的示唆

Kimi K3の重み公開は、単なる「非常に大きな中国製LLMの公開」ではない。研究面では、KDA、AttnRes、極端に疎なLatentMoE、ネイティブ視覚、100万トークン、エージェントRLを2.8兆パラメータ規模で統合した実用モデルを外部が直接調査できるようにした点が重要である。産業面では、閉鎖型フロンティアモデルに近い能力を、APIだけでなく自己運用可能な重みとして配布することで、推論エンジン、クラウド、量子化、エージェント基盤を含む市場構造に圧力を加えた。⁵⁹

一方で、「2.8兆パラメータ」「オープン」という言葉だけから、低コスト、完全透明、再現可能、安全、無条件商用利用可能と結論付けることはできない。実際には、64基以上のアクセラレーターが推奨され、学習データと計算量は非公開で、独自ライセンスには大規模MaaS条件があり、安全性・バイアス文書も不足している。⁶⁰

組織別の推奨対応

対象	推奨対応	理由
研究機関	重み、設定、技術報告書の版を固定し、チェックサム、推論エンジン、ハーネス、推論予算を記録する	公式表はハーネスが混在し、将来更新され得る
ベンチマーク研究者	公式のKimi Code条件だけでなく、共通ハーネス、同一GPU、同一トークン予算で再評価する	1〜3ポイント差はハーネス差の可能性がある
一般企業	まずAPIまたは認定ホストで限定PoCを行い、セルフホストはTCO比較後に判断する	64基以上の推奨構成と約1.5TBの重みが必要
機密データ利用企業	APIのデータ保持・学習利用・越境移転を確認し、セルフホスト時もログとRAG基盤を暗号化する	外部送信を止めても内部管理責任は残る
SaaS企業	MAU、月次売上、関連会社を含む12か月売上を法務部門で判定する	MaaSの2000万ドル条件と大規模サービス表示義務がある
EU向け事業者	GPAI提供者該当性、オープンソース免除、EU代理人、学習内容要約、著作権方針を確認する	2026年8月2日から完全執行が始まる

対象	推奨対応	理由
セキュリティ責任者	サンドボックス、最小権限、ネットワーク分離、秘密情報マスキング、監査ログ、実行承認を必須化する	サイバー評価で攻撃支援試行を安全策が止めなかった
調達・広報担当	「オープンソース」ではなく「独自ライセンスのオープンウェイト」と表記する	データ・学習過程の非公開と商用条件がある

総合判断

研究価値：高い。 3T級の疎モデル、線形アテンション、ネイティブマルチモーダル、長時間エージェントの実モデルを調査できることは、公開モデル研究にとって大きな前進である。

性能価値：高いが領域依存。 コーディング、ブラウジング、ツール利用、文書・視覚処理では最上位級だが、難問推論ではClaude Fable 5やGPT-5.6 Solへの劣後が残る。日本語特化性能は未公開である。 ²²

運用価値：大規模事業者には高いが、一般利用者には限定的。 有効パラメータは1040億に抑えられていても、重み全体の保管と専門家並列通信が必要であり、通常のワークステーションでの実用運用は想定しにくい。 ⁶¹

透明性：中程度。 重みとアーキテクチャは公開されたが、学習データ、計算量、包括的安全性評価は不足している。

法的確実性：中～低。 通常の商用利用は可能だが、独自のMaaS売上条件、EU規制、学習データ不透明性、米中の蒸留・輸出管理をめぐる未解決の争点がある。 ⁶²

結論として、Kimi K3の重み公開は、**2026年のオープンウェイトAIにおける技術的・産業的な重要事件**である。ただし、その価値は「誰でも自由かつ容易に使える完全オープンなモデル」という点ではなく、**フロンティア級のエージェントモデルを外部が保有・検査・変更できるようになった点**にある。採用判断では、性能表だけでなく、インフラ総費用、独自ライセンス、学習データ透明性、安全制御、地政学的供給リスクを同じ重みで評価する必要がある。

¹ ⁴ ⁷ ¹⁰ ¹⁵ ²⁸ ⁵⁰ ⁶⁰ <https://www.kimi.com/blog/kimi-k3>
<https://www.kimi.com/blog/kimi-k3>

² ³ ¹³ ¹⁴ ²¹ ²² ²³ ⁴⁵ ⁶¹ <https://huggingface.co/moonshotai/Kimi-K3>
<https://huggingface.co/moonshotai/Kimi-K3>

⁵ ¹¹ ⁶² <https://github.com/MoonshotAI/Kimi-K3/blob/main/LICENSE>
<https://github.com/MoonshotAI/Kimi-K3/blob/main/LICENSE>

⁶ ¹⁸ ¹⁹ ²⁵ <https://arxiv.org/pdf/2607.24653>
<https://arxiv.org/pdf/2607.24653>

⁸ ²⁰ <https://github.com/MoonshotAI/Kimi-K3>
<https://github.com/MoonshotAI/Kimi-K3>

⁹ ³⁶ ⁵⁹ <https://arxiv.org/abs/2607.24653>
<https://arxiv.org/abs/2607.24653>

- 12 <https://www.nist.gov/news-events/news/2026/07/uk-aisi-caisi-preliminary-assessment-kimi-k3s-cyber-capabilities>
<https://www.nist.gov/news-events/news/2026/07/uk-aisi-caisi-preliminary-assessment-kimi-k3s-cyber-capabilities>
- 16 <https://sebastianraschka.com/blog/2026/kimi-k3-architecture-notes.html>
<https://sebastianraschka.com/blog/2026/kimi-k3-architecture-notes.html>
- 17 <https://arxiv.org/abs/2510.26692>
<https://arxiv.org/abs/2510.26692>
- 24 <https://artificialanalysis.ai/articles/kimi-k3-agentic-knowledge-benchmark>
<https://artificialanalysis.ai/articles/kimi-k3-agentic-knowledge-benchmark>
- 26 <https://www.aisi.gov.uk/blog/preliminary-assessment-of-kimi-k3s-cyber-capabilities>
<https://www.aisi.gov.uk/blog/preliminary-assessment-of-kimi-k3s-cyber-capabilities>
- 27 <https://news.ycombinator.com/item?id=49085698>
<https://news.ycombinator.com/item?id=49085698>
- 29 39 <https://www.reuters.com/world/china/chinas-moonshot-unveils-worlds-largest-open-ai-model-closing-us-rivals-2026-07-17/>
<https://www.reuters.com/world/china/chinas-moonshot-unveils-worlds-largest-open-ai-model-closing-us-rivals-2026-07-17/>
- 30 <https://www.axios.com/2026/07/18/kimi-k3-ai-models-china-us-japan>
<https://www.axios.com/2026/07/18/kimi-k3-ai-models-china-us-japan>
- 31 <https://www.reuters.com/legal/transactional/chinas-moonshot-pauses-kimi-subscriptions-amid-hot-demand-ipo-push-2026-07-20/>
<https://www.reuters.com/legal/transactional/chinas-moonshot-pauses-kimi-subscriptions-amid-hot-demand-ipo-push-2026-07-20/>
- 32 <https://techcrunch.com/2026/07/20/openai-is-scared-of-open-weight-models-should-the-us-be/>
<https://techcrunch.com/2026/07/20/openai-is-scared-of-open-weight-models-should-the-us-be/>
- 33 <https://vllm.ai/blog/2026-07-22-kimi-k3-preview>
<https://vllm.ai/blog/2026-07-22-kimi-k3-preview>
- 34 41 53 <https://www.reuters.com/legal/litigation/ai-grows-more-powerful-us-china-feud-threatens-safety-efforts-2026-07-24/>
<https://www.reuters.com/legal/litigation/ai-grows-more-powerful-us-china-feud-threatens-safety-efforts-2026-07-24/>
- 35 43 https://x.com/Kimi_Moonshot/status/2081760186235289764
https://x.com/Kimi_Moonshot/status/2081760186235289764
- 37 <https://zenn.dev/upgradetech/articles/c129ae0ecd3cd7>
<https://zenn.dev/upgradetech/articles/c129ae0ecd3cd7>
- 38 <https://times.abema.tv/articles/-/10264370>
<https://times.abema.tv/articles/-/10264370>
- 40 <https://www.theverge.com/ai-artificial-intelligence/971444/how-chinese-open-weight-ai-models-impact-us-companies>
<https://www.theverge.com/ai-artificial-intelligence/971444/how-chinese-open-weight-ai-models-impact-us-companies>
- 42 49 <https://exawizards.com/column/ai-trend/news-07-28-2026-3/>
<https://exawizards.com/column/ai-trend/news-07-28-2026-3/>
- 44 https://www.reddit.com/r/LocalLLaMA/comments/1uy3a0q/kimi_k3_released_on_web_and_app/
https://www.reddit.com/r/LocalLLaMA/comments/1uy3a0q/kimi_k3_released_on_web_and_app/

46 <https://huggingface.co/moonshotai>

<https://huggingface.co/moonshotai>

47 48 https://github.com/MoonshotAI/Kimi-K3/blob/main/k3_tech_report.pdf

https://github.com/MoonshotAI/Kimi-K3/blob/main/k3_tech_report.pdf

51 <https://opensource.org/blog/state-of-the-source-at-ato-2025-state-of-the-open-ai>

<https://opensource.org/blog/state-of-the-source-at-ato-2025-state-of-the-open-ai>

52 <https://opensource.org/osd>

<https://opensource.org/osd>

54 <https://www.reuters.com/world/china/china-accuses-us-ai-hegemonism-threatens-countermeasures-over-potential-probes-2026-07-27/>

<https://www.reuters.com/world/china/china-accuses-us-ai-hegemonism-threatens-countermeasures-over-potential-probes-2026-07-27/>

55 56 <https://digital-strategy.ec.europa.eu/en/faqs/guidelines-obligations-general-purpose-ai-providers>

<https://digital-strategy.ec.europa.eu/en/faqs/guidelines-obligations-general-purpose-ai-providers>

57 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20260331_1.pdf

https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20260331_1.pdf

58 https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/

https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/