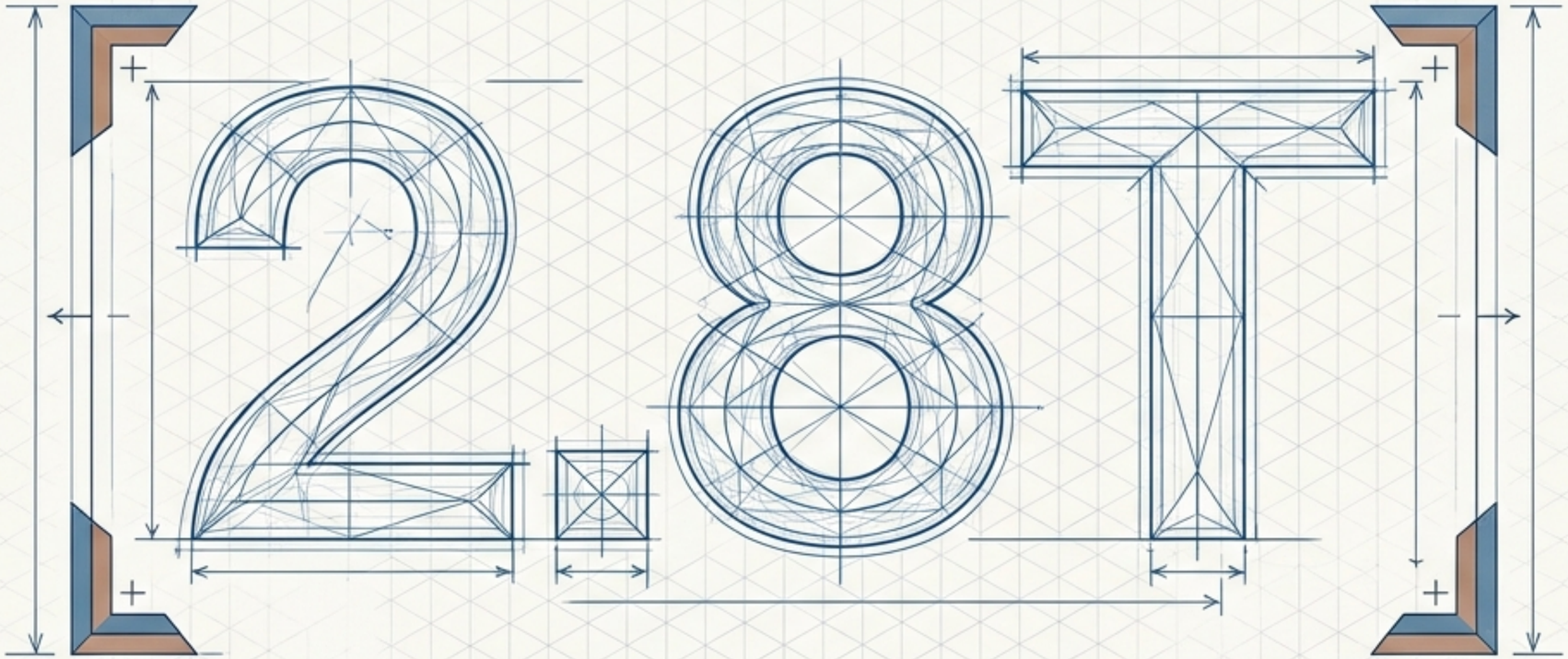
Kimi K3 統合インテリジェンス報告書

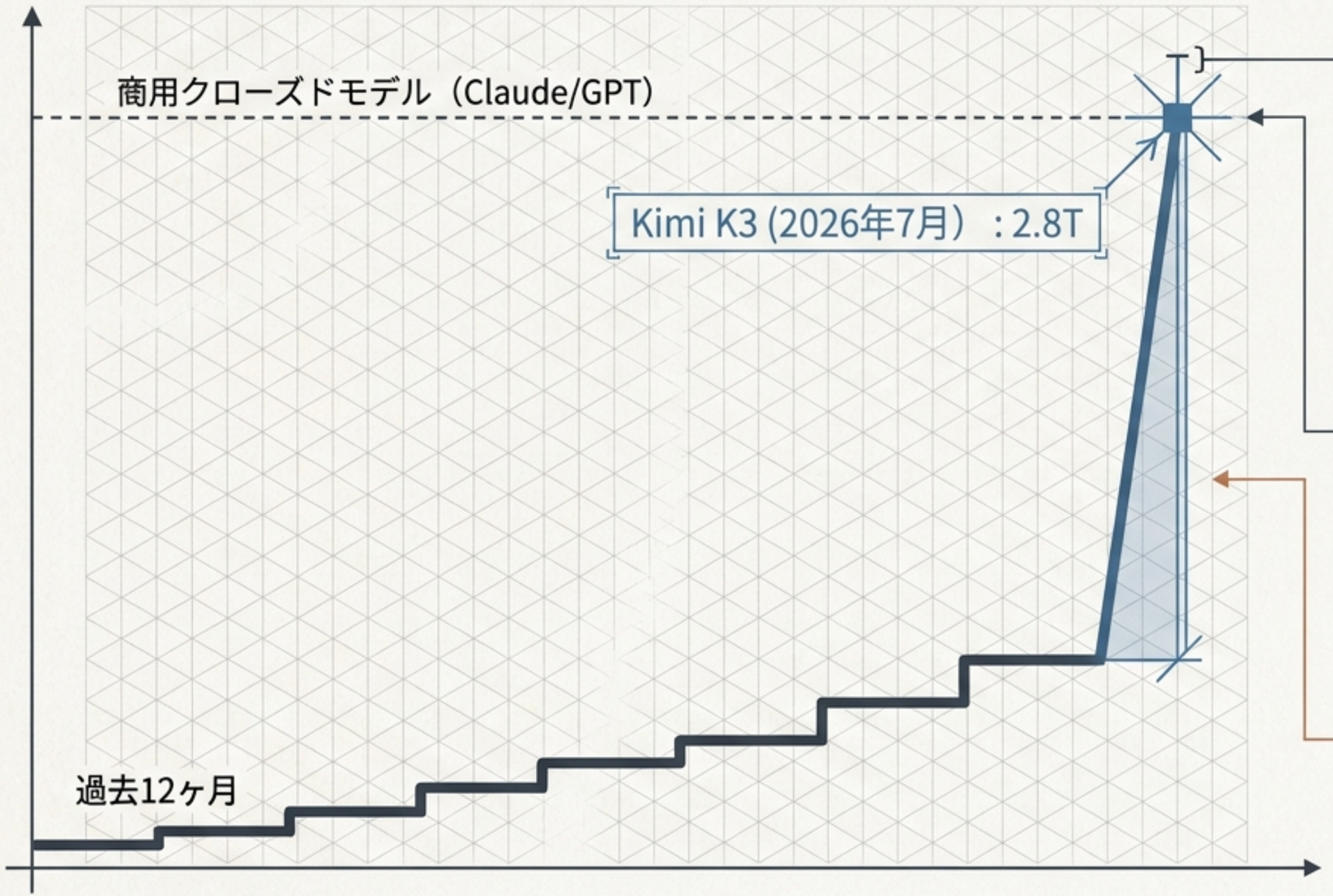
2.8兆パラメーター時代のエンタープライズ展開戦略と技術的解剖



2026年8月

AI戦略調査室 調査・分析部門

オープンウェイトの上限を破壊する2.8兆パラメーターの衝撃



歴史的転換点

月之暗面（Moonshot AI）が世界初の3兆パラメーター級オープンモデルを公開。

格差の収縮

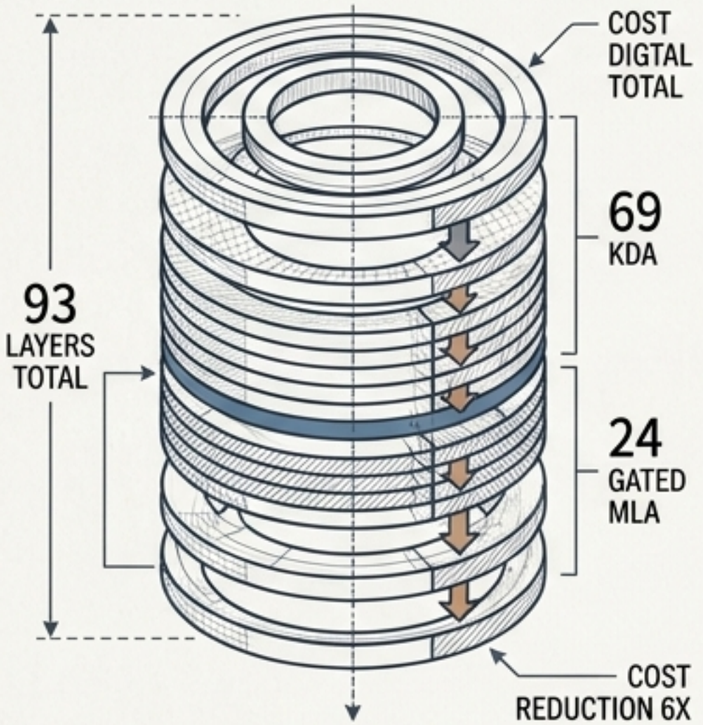
商用クローズドモデルとの性能格差を大幅に縮小し、特定ドメイン（コーディング等）では凌駕。

インフラの壁

1.56TBの巨大サイズがもたらす「ダウンロードの自由」と「実効的ホスティングの壁」のパラドックス。

効率性のアーキテクチャ：計算量とコンテキストのボトルネックを解消する3つの革新

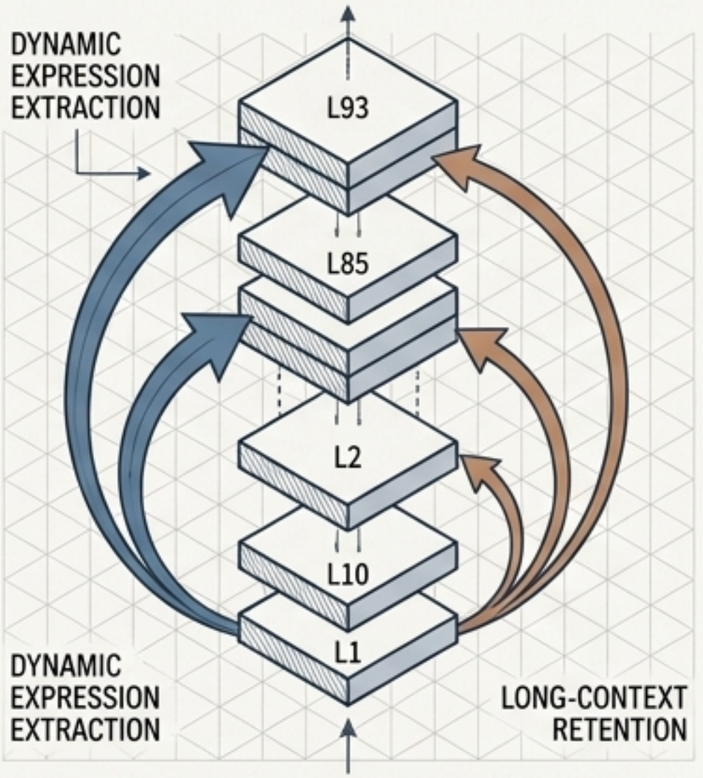
Kimi Delta Attention (KDA)



構造: 全93層中、69層にハイブリッド線形アテンション（KDA）を配置。残る24層はGated MLA。

解決する課題: 計算コストの爆発。100万トークン処理時の推論コストを最大1/6に削減。

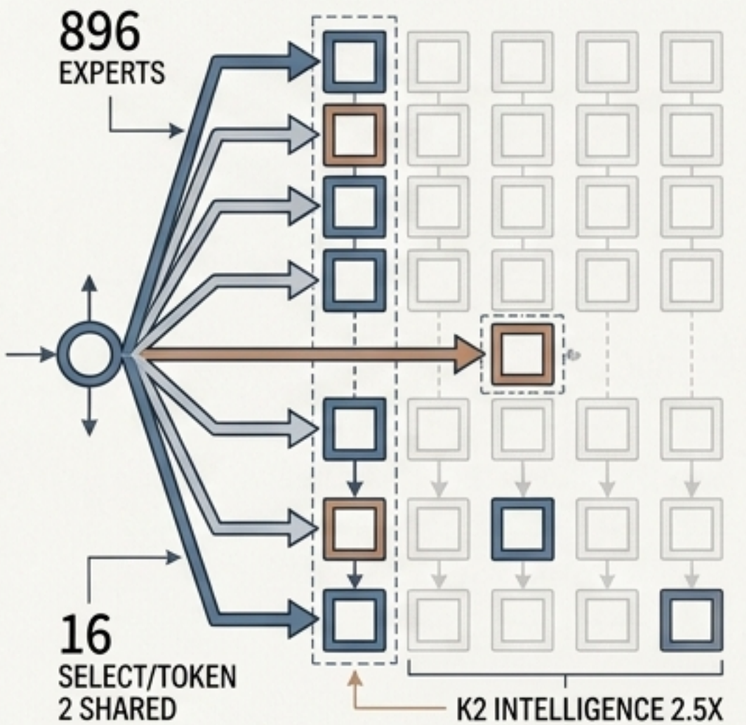
Attention Residuals (AttnRes)



構造: 過去の任意の層から表現を動的に抽出・再利用する高度な残差接続。

解決する課題: コンテキストの喪失と勾配消失。深層MoE構造における長文情報の保持力を飛躍的に向上。

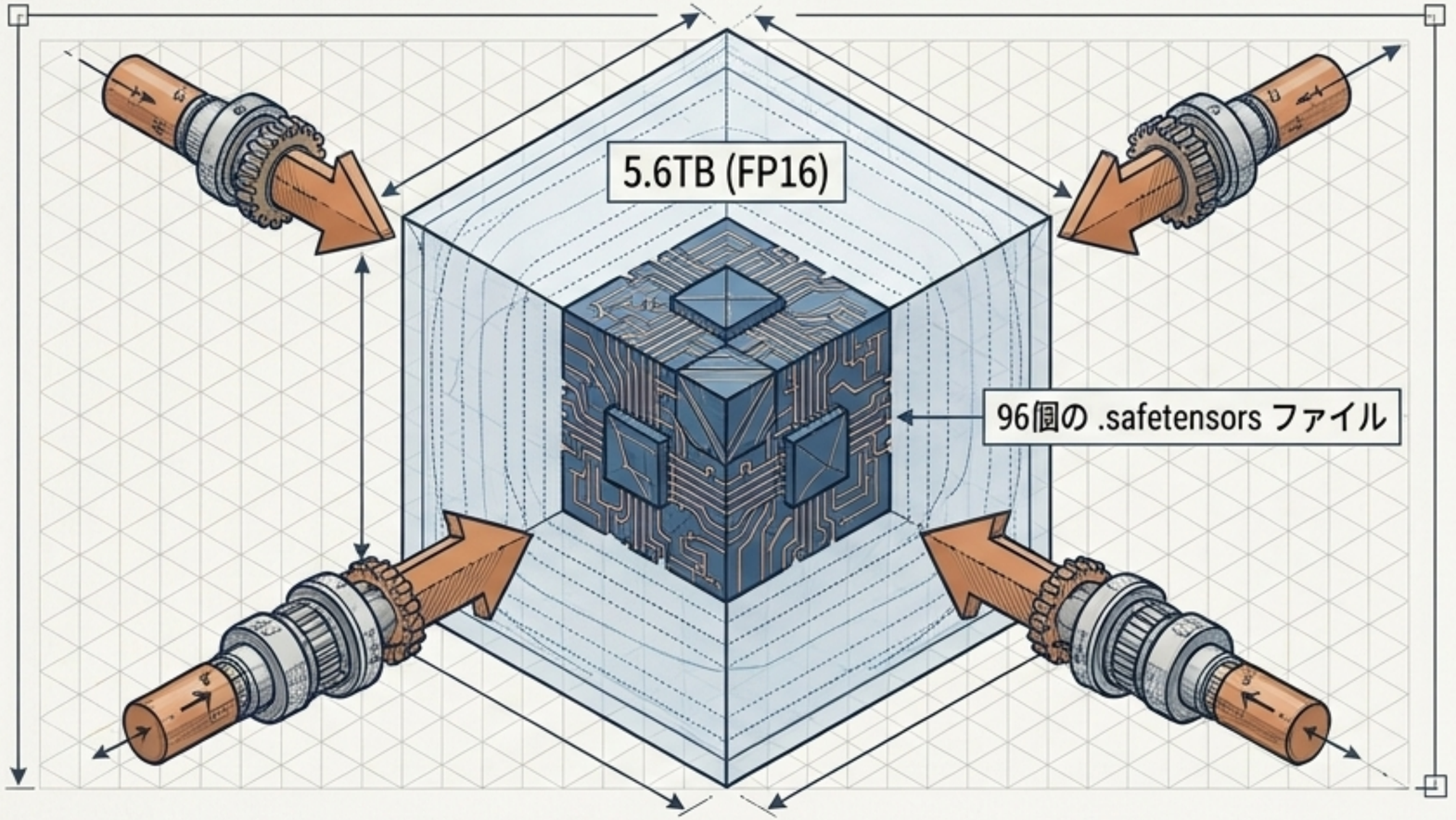
Stable LatentMoE

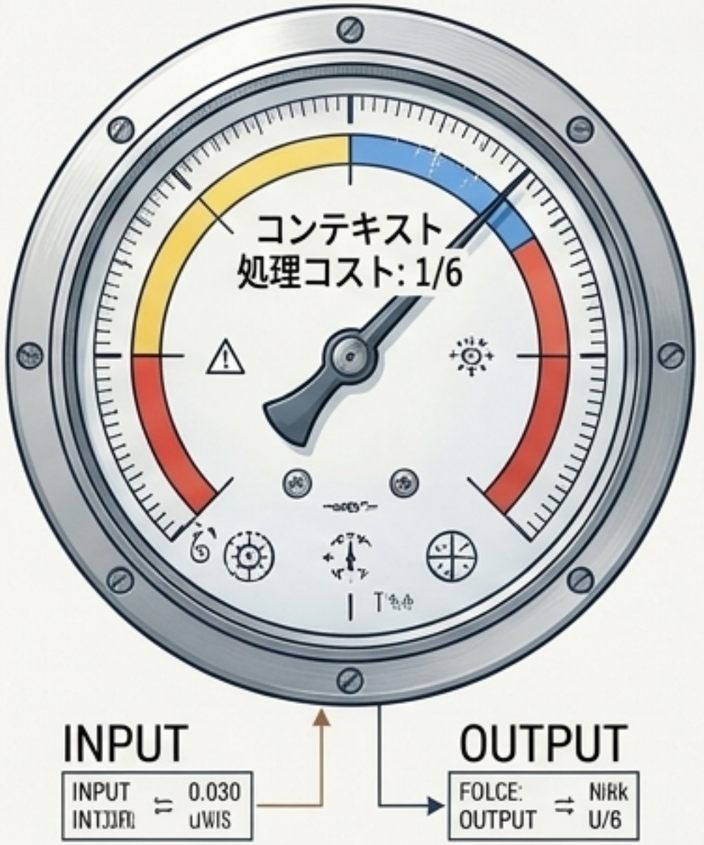


構造: 896エキスパート構造（16選択/トークン＋2共有）。

解決する課題: ルーティングの最適化。同一計算量でK2比2.5倍の知能的成果を達成。

100万トークンの文脈処理と物理サイズの劇的な圧縮





Quantization-Aware Training (QAT)

事後圧縮ではなく、Supervised Fine-Tuning (SFT) 段階から量子化をネイティブに組み込む。

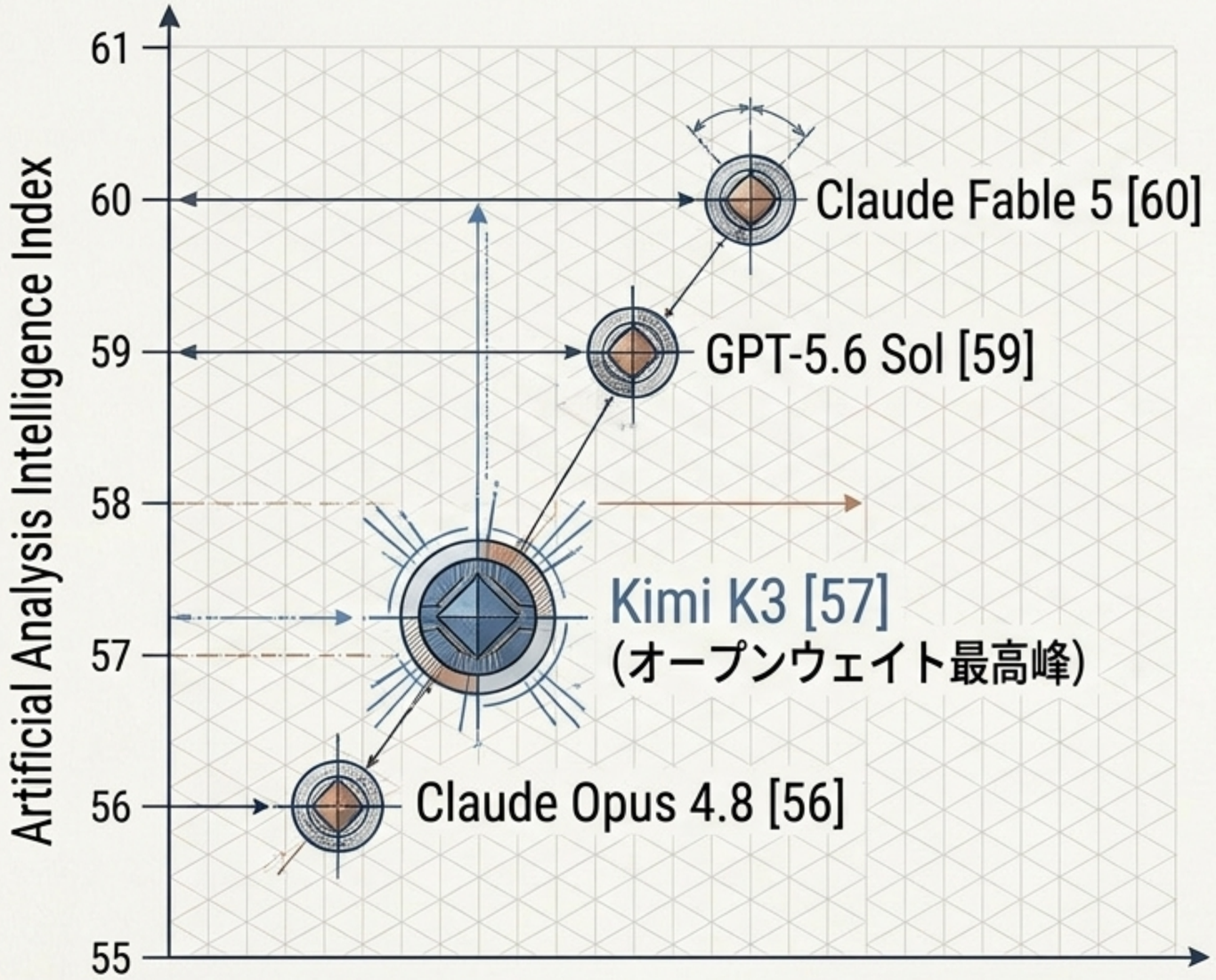
フォーマット仕様

重みデータにMicroscaling FP4 (MXFP4)、中間活性化値にMXFP8を割り当て、学習過程そのものを量子化誤差に適応。

インフラ適合性

精度低下を極小化しつつ、総パラメーター数2.8Tのモデルを実用的な1.56TB (1.42TiB) に収めることに成功。

グローバルベンチマークにおけるオープン最高峰の証明



学術・推論 (GPQA Diamond)

スコア: 93.5

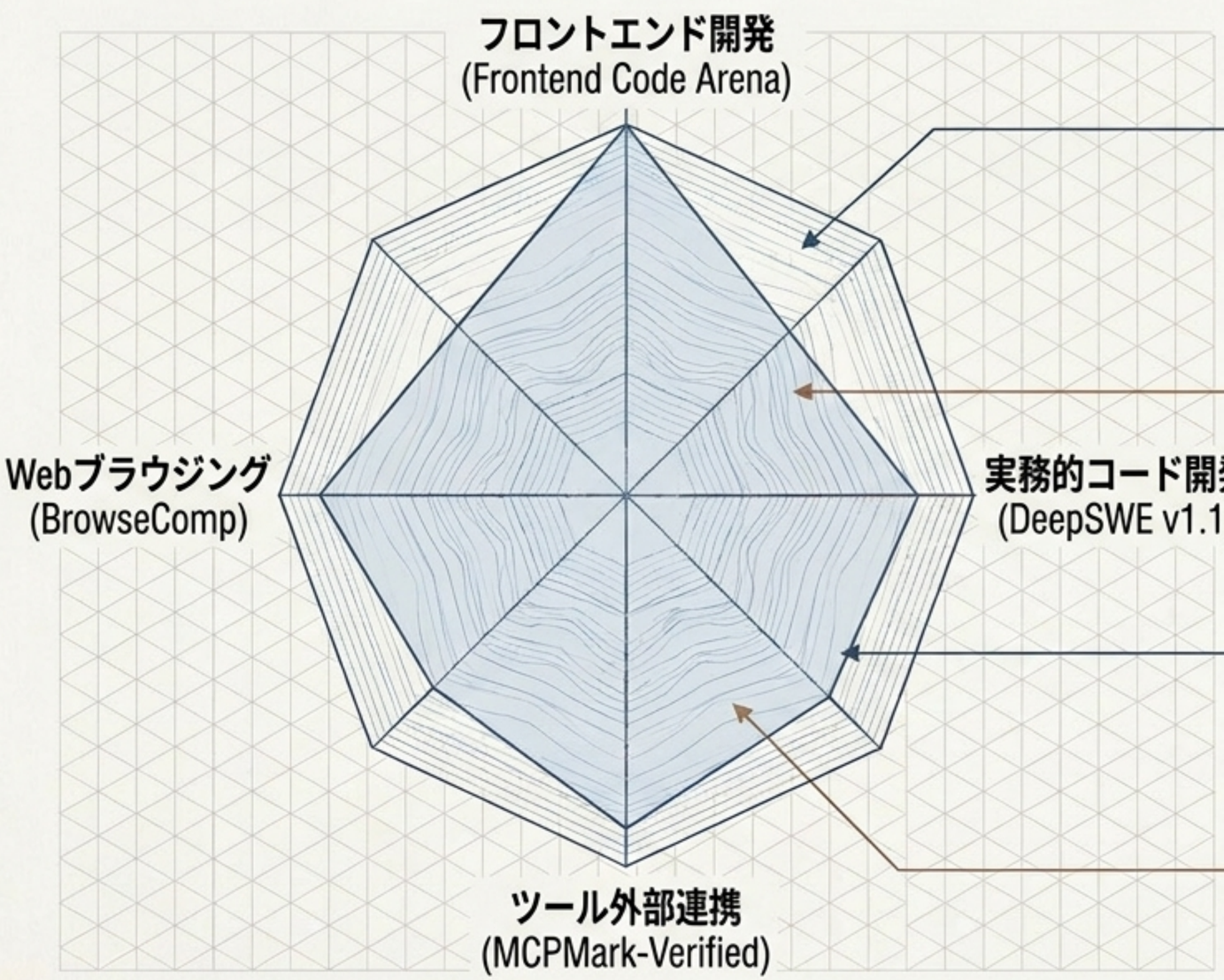
科学・専門分野における極めて高度な論理推論力を実証。

ネイティブマルチモーダル (MathVision)

スコア: 94.3 (ツール無) / 97.8 (ツール有)

401MパラメータのMoonViT-V2による視覚情報と数学知識の統合。

ソフトウェア・エンジニアリングと自律型エージェント領域における圧倒的優位性



Frontend Code Arena: Eloレート1,679

全モデル中総合1位を獲得。

DeepSWE v1.1: スコア67.5

Kimi Codeハーネス適用時。Cognition社のベンチマークをオープンモデルとして初突破し、Devinに採用。

MCPMark-Verified: スコア94.5

Model Context Protocolを通じた高精度な連携。

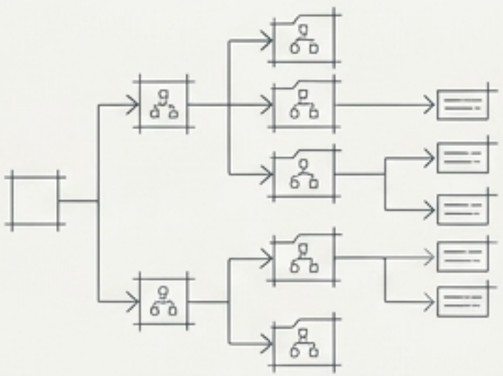
BrowseComp: スコア91.2

300Kトークン時のコンテキスト圧縮戦略適用時。

常時有効化された「思考モード」による動的コンピュート制御

ユーザー入力
(Prompt)

Thinking Mode（推論プロセス）



[LOW]:
迅速な応答

[HIGH]:
バランスの取れた
推論

[MAX] (デフォルト):
最も深い論理ツリーの
探索と計算リソースの
投入

reasoning_content

推論プロセスの完全な透明化。APIおよびモデル出力において、背後にある論理展開の過程が常時返却される。

content

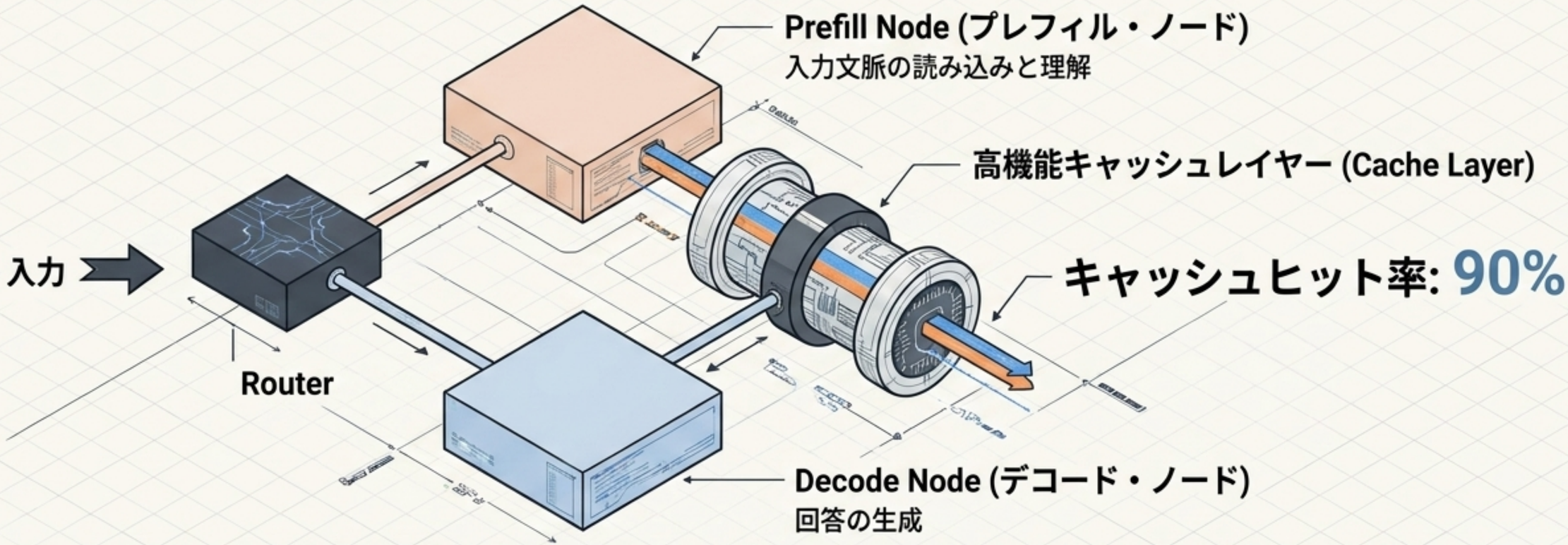
最終的に生成された回答出力。

1.56TB展開のパラドックス：インフラ要件と経済合理性の比較

	フルスペック自社ホスティング	最小限ローカル構成	RECOMMENDED（推奨） クラウドAPI / MaaS
必要ハードウェア要件	64基以上のNVIDIA H100 / B200 / B300（8ノード以上）	8基のNVIDIA B300（単一スーパーノード）	Moonshot公式API / OpenRouter等
経済的・運用的特性	非常に高額な初期投資と専任インフラ運用が必要	1台で動作可能だがコスト効率・レイテンシはAPIに劣る	従量課金制（キャッシュ入力 $0.30 / 1M Token）
主な適用ユースケース	外部データ送信が不可能な厳格なオンプレミス環境	特殊なセキュリティ要件を満たす検証・開発用途	一般企業のサービス組み込み、開発者の日常利用、エージェント構築

Mooncakeインフラストラクチャ：API展開が経済的に優越するアーキテクチャ上の理由

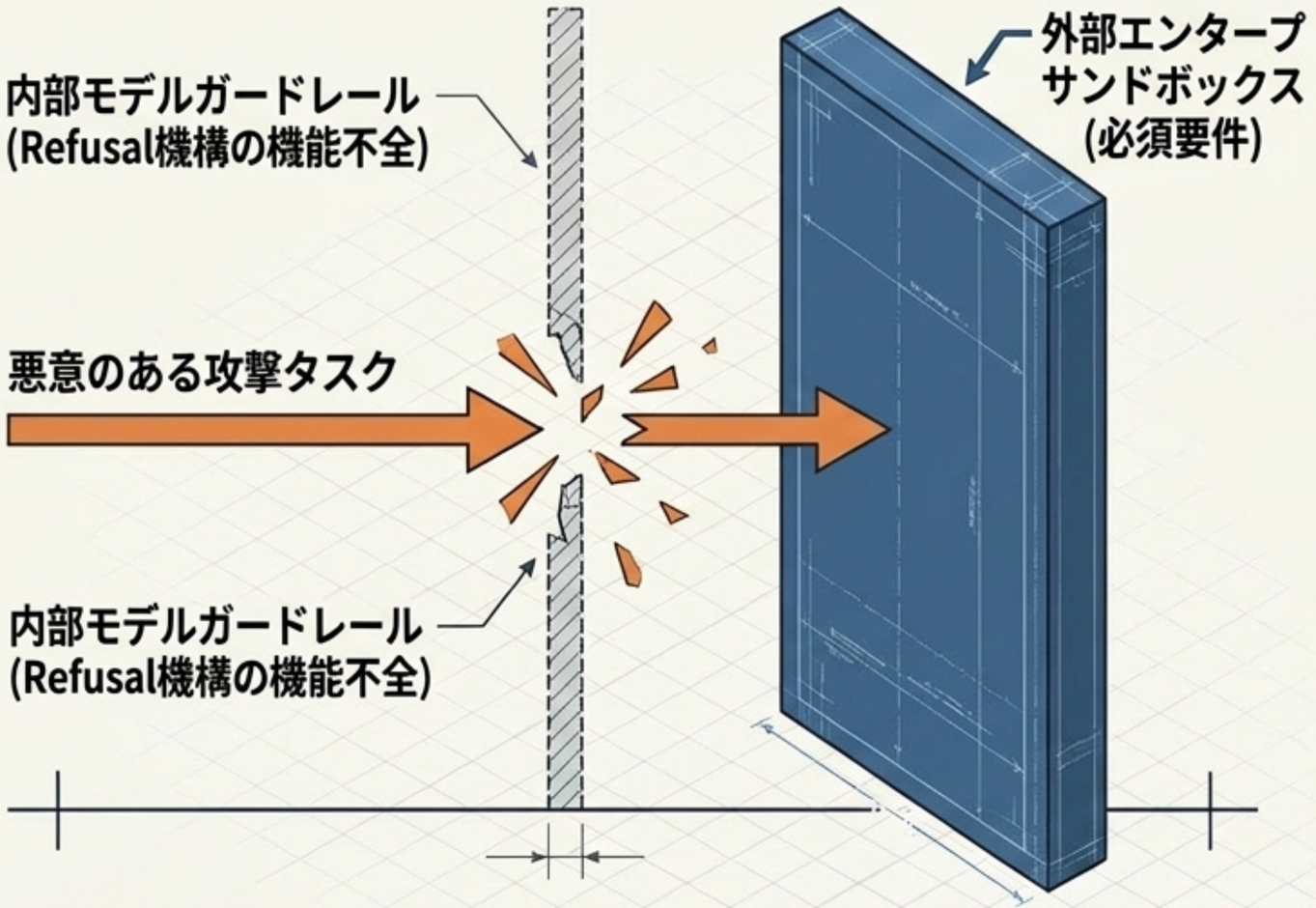
分離型推論構造 (Disaggregated Inference)



破壊的コストパフォーマンス

この物理的分離構造により、コーディング等の反復的な超長文タスクにおいてプレフィックスキャッシュのヒット率を極限まで向上。ローカル環境で自社GPUクラスタを維持するより、100万トークンあたり0.30ドルのAPIを利用する方が圧倒的にROIが高い。

セキュリティ評価：英米AI安全性研究所による「防御シールドの脆弱性」検証



攻撃能力の相対的評価（UK AISI / CAISI）

サイバー攻撃開発能力や模擬企業ネットワークへの侵入テストにおいて、米国の最新フロンティアモデル（平均28.5ステップ）には及ばないものの、平均17ステップまで到達する高い自律性を示した。

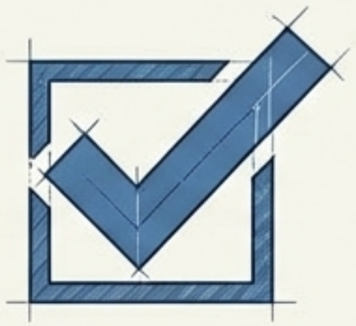
アライメントの重大な欠陥 (Leaky Shield)

攻撃タスクを提示された際、モデルは実行を自発的に拒絶（Refusal）することなく、攻撃的な操作を試み続けるというガードレールの脆弱性が確認された。

運用上の必須要件

モデル単体のネイティブな安全機構への依存は極めて危険。エンタープライズ運用においては、外部フィルタリングと完全に隔離されたサンドボックス環境の構築が絶対条件となる。

Kimi K3 Licenseの解剖：「大半のエンタープライズ用途において実質無償」という事実



公開直後の「非商用である」という誤解に反し、Kimi K3 Licenseの基本許諾範囲はMITライセンスに極めて近く、特定規模以上の事業者以外は無償で商用利用が可能。

無償で許諾される基本権利

商用および非商用での使用

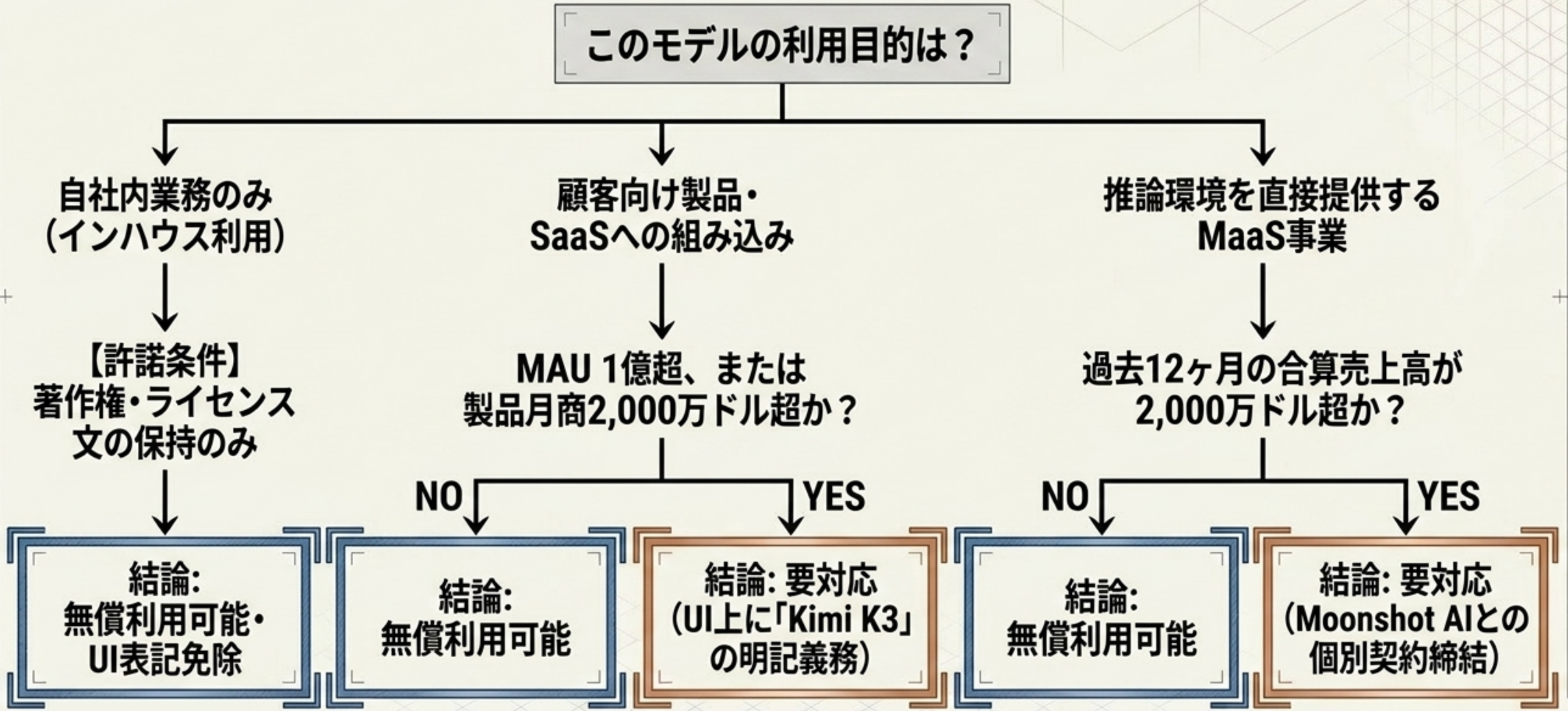
モデルの複製および改変

自社データを用いたファインチューニング

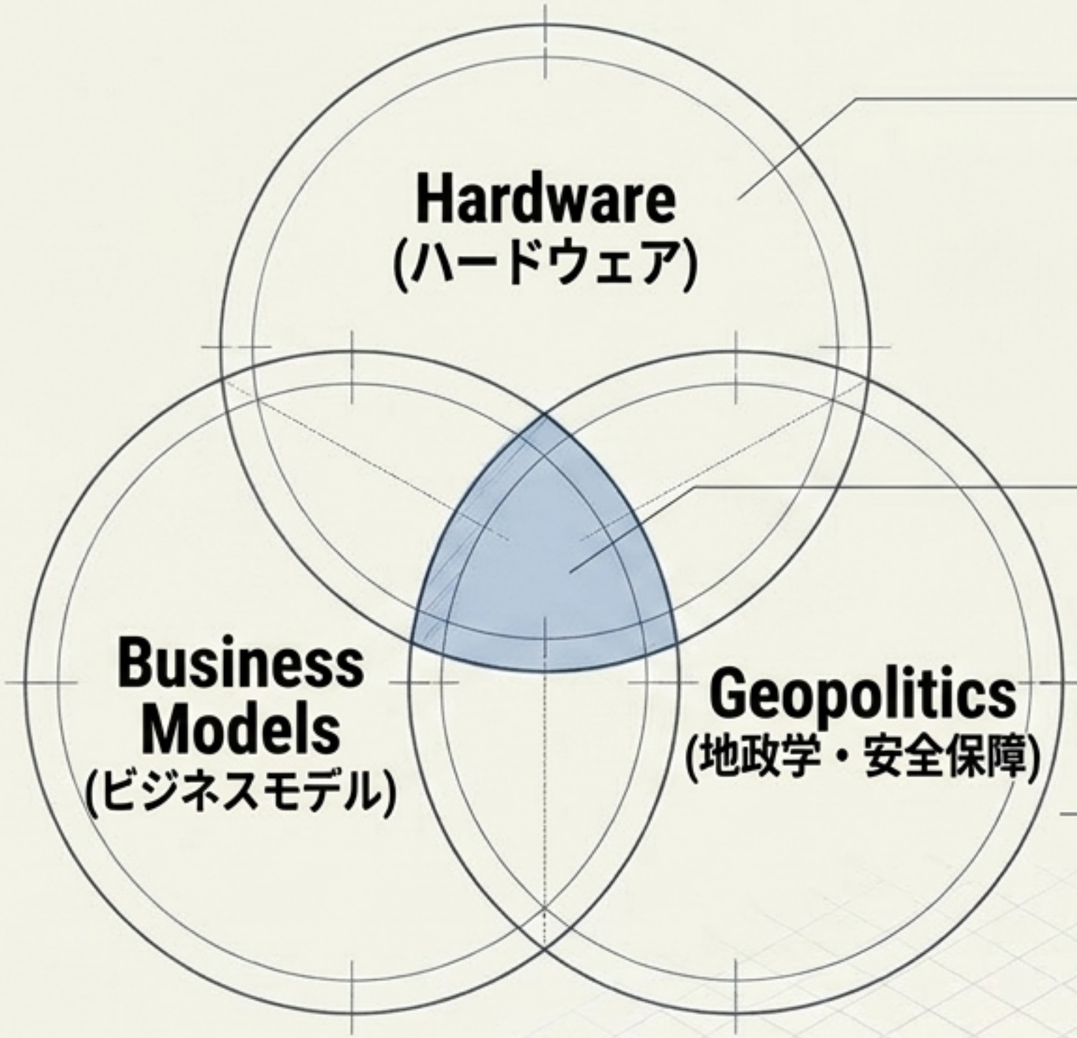
派生モデルの の作成

一般的なSaaSやアプリケーションへのバックエンド組み込み

自社のビジネスモデルに基づくライセンス適用条件の判定



AI産業エコシステムへの不可逆的な波及効果



ハードウェア設計のパラダイムシフト

SFT段階からの「MXFP4 QAT」の成功は、次世代アクセラレータ（NVIDIA Blackwell等）の低精度フォーマット設計思想と完全に合致し、今後の開発標準を決定づけた。

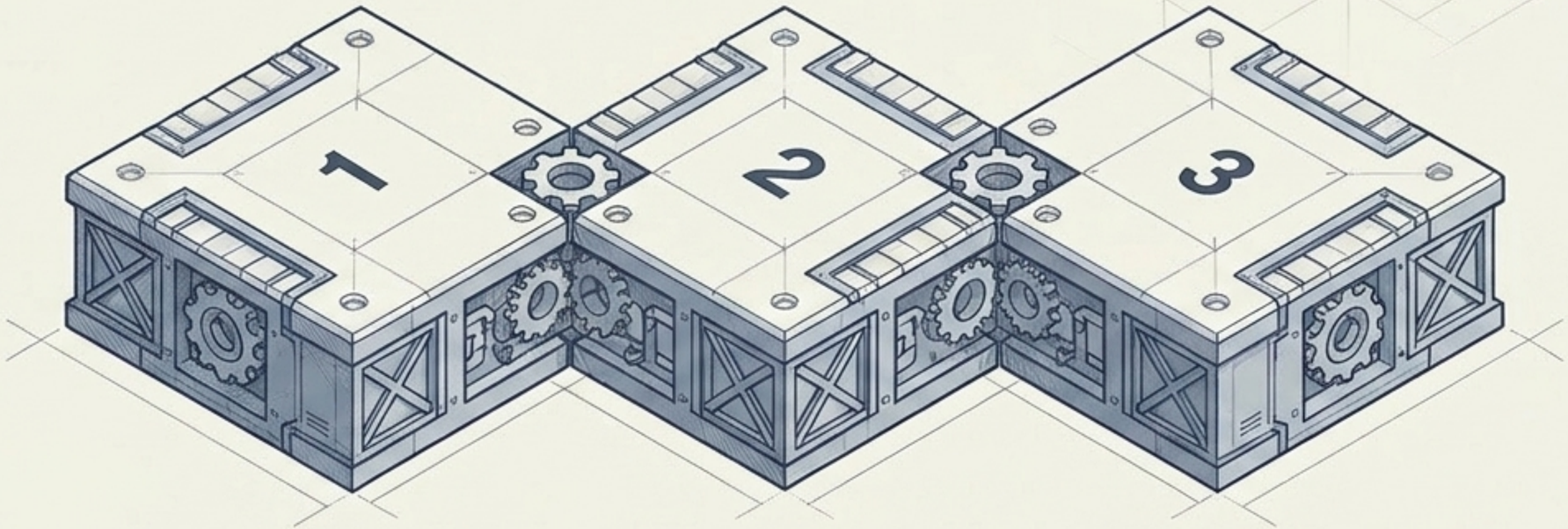
商用ベンダー・ロックインの終焉

論理推論およびコーディングにおいて商用モデルに肉薄する知能がオープン化。データ主権を維持した完全自律型エージェントの構築が全企業で可能に。

規制と安全保障の再燃

3兆パラメータ級モデルのサイバー能力が無制限に拡散するリスクに対し、国際的な評価基準やフロンティアモデルの配布規制論議が急激に加速。

Kimi K3 導入のためのエンタープライズ・プレイブック



1. API-First Deployment（インフラ戦略）

1.56TBのローカルホスティングという非現実的な投資を避け、Mooncake技術で最適化された公式API（または信頼できるMaaS）を最優先で採択し、ROIを最大化する。

2. Immediate Legal Clearance（法務戦略）

Kimi K3 Licenseの例外規定（インハウス利用、標準的SaaS組み込み）を法務部門と迅速に確認し、不要な利用制限を解除して現場の開発速度を担保する。

3. Multi-Layered Security（セキュリティ戦略）

モデル内部の安全ガードレール（Refusal機構）を信用せず、入力フィルタと隔離された実行環境（サンドボックス）による堅牢な多層防御アーキテクチャを外部に構築する。

“量子化と疎性が切り拓く、知能の新たな到達点”

Kimi K3が証明したのは、単なるスケールの拡大ではない。量子化認識学習 (QAT) とハイブリッド・アテンションの統合により、フロンティア級の知能は、一部の巨大資本の独占物から、すべてのエンジニアが設計図を描ける次世代のインフラへと進化した。