

Mistral Large 4 徹底解剖： 1兆パラメータ・モデルの技 術的現実

エンタープライズ運用のためのハードウェア要件、
制約、および導入前評価見取り図

2026年10月 | 独立技術調査委員会

アーキテクチャ (Granular MoE)

1.05T

Total Parameters

49B

Active per Token
(約4.67%)

入出力能力

1M Context Window

1.6B Vision Encoder
(Text + Image Input)

学習環境



3,800 Grace
Blackwell GPUs

欧州自社DC / 160+言語

公開ステータス

Public Preview

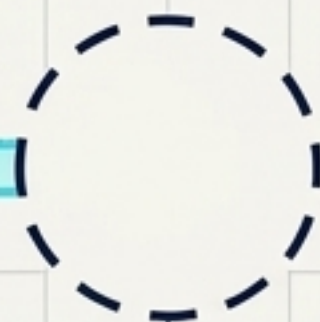
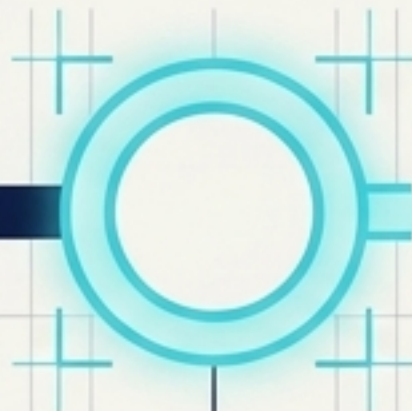
2026年10月6日発表・API提供中

開発ロードマップ

2026-10-06

Current Phase

2026-10月末 (Target)



Public Preview 開始

API (Mistral Studio)
経由での提供開始。
Le Chonk 発表。

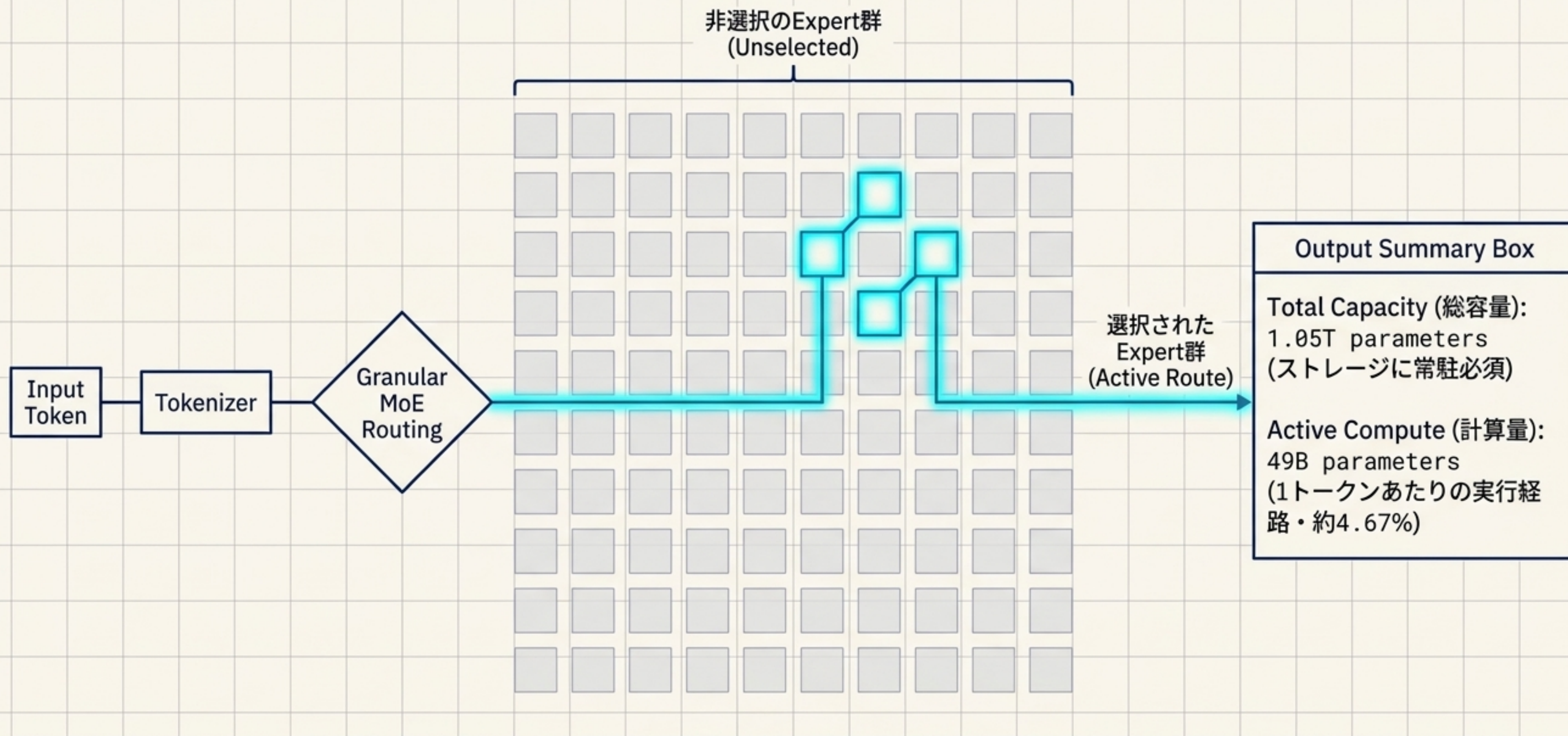
Red-Teaming & 評価期間

サイバーセキュリティ企業・
政府機関によるテスト。
RLトレーニングの継続。

Open-Weight 公開予定

Reuters等は10月27日と報道。
公式は月末まで。
詳細アーキテクチャ開示。

現状のステータス: 現時点ではAPI経由のプレビューのみ。
ウェイトファイル、フォーマット、正式なライセンスは未確定。



注意: 49Bモデルと同じGPUで動くという解釈は誤り。計算量は少ないが、巨大な重みデータを保持するVRAM領域が必須。



100万トークンの物理的重さ

- **KVキャッシュの膨張**: 128kから1Mでトークン数は約7.8倍に増加し、メモリを圧迫する。
- **不確実性**: ML4のレイヤー数、KVヘッド数、アテンション方式が未公表のため、1Mフル稼働時の正確なVRAM消費量は現在算出不可。
- **結論**: 1M仕様は存在するが、APIでの実効性能とコストには依然として制約が見られる。

Known / Unknown Matrix

確定済み情報 (Known Realities)

学習インフラ:

3,800 Grace Blackwell GPUs
(欧州自社DC)



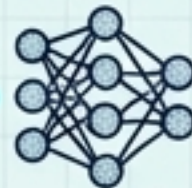
対応言語:

160以上の言語 (全EU公用語を含む)



パラメータ:

Total 1.05T / Active 49B



安全性:

B3 Attack Resistance 93.3%



未公表・ブラックボックス (The Unknowns)

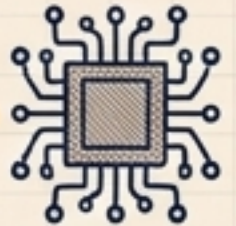
データ透明性:

事前学習の総トークン数、
データセット構成



アーキテクチャ詳細:

トークナイザ方式、Expert数、
Attention方式



ライセンス:

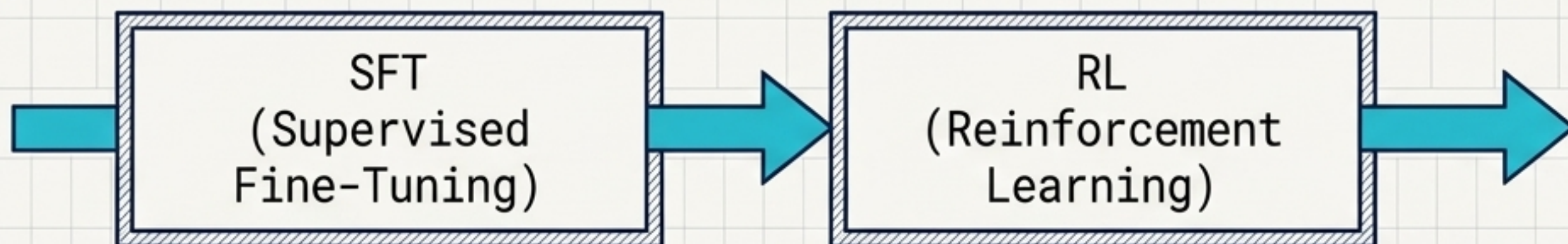
ウェイト公開時の商用利用条件
(Apache-2.0と断定不可)



総括: 強力なインフラを証明したが、完全なオープン性には達していない。
技術的透明性は月末のウェイト公開に依存する。

POST-TRAINING PIPELINE OVERVIEW

FACTORY PIPELINE



約3,000 GPUで1日約**330億**トークンを生成
→ フィルタリング後、約**160億**の学習用トークンを抽出。

SAFETY ALIGNMENT ENGINE



高度なサイバー推論能力の維持と**悪意あるリクエストへの高い拒否率**の両立。

運用上のインサイト

モデル内アライメントは確立されている。

ただし、セルフホスト導入時はAPIゲートウェイ層での独自のアクセス制御やサンドボックス化が不可欠。

SOTA級の専門領域

Cybersecurity

Cybench 93% / B3 Attack Resistance 93.3%



Agentic Coding

DeepSWE 1.1 61.7%



Enterprise Workflow

AA-Briefcase 1,393 Elo

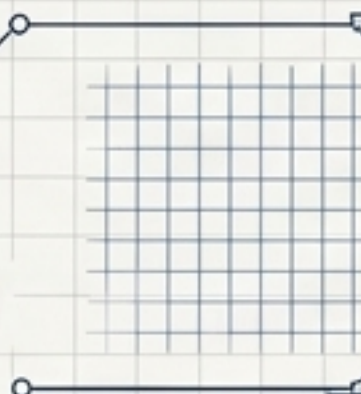


汎用性能

総合知能指数



AA Intelligence Index 38点 (225モデル中64位)



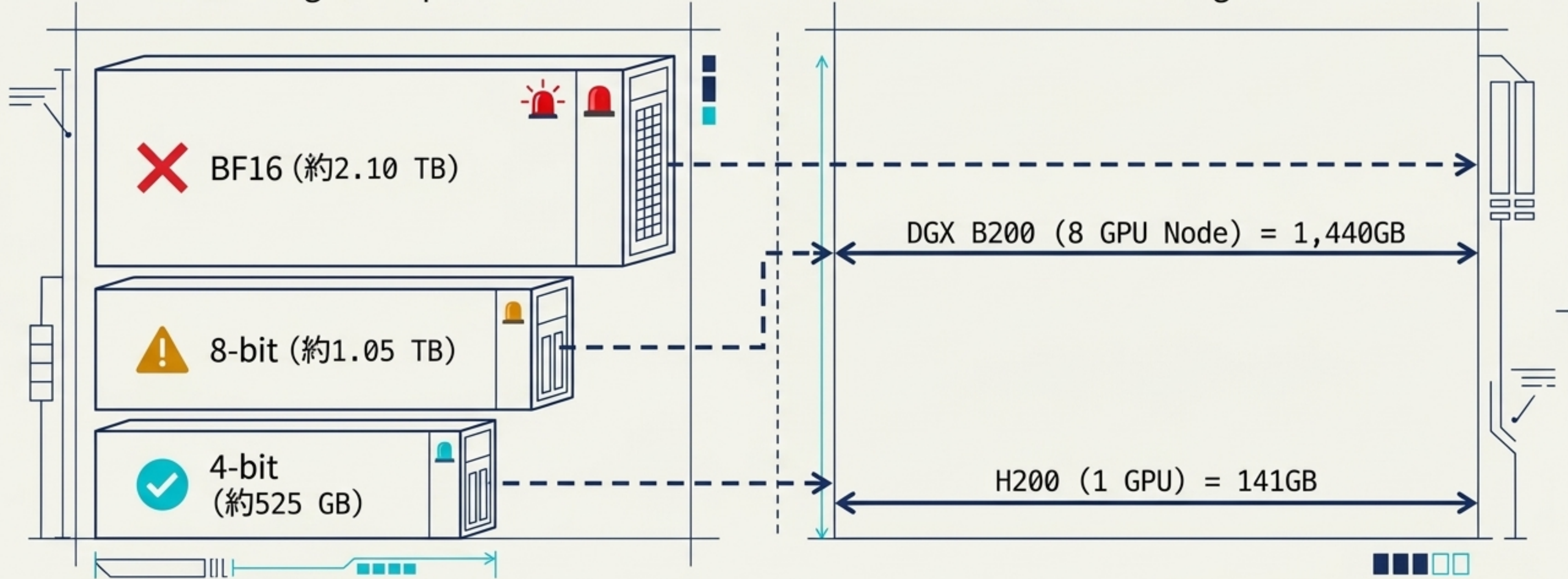
万能な汎用SOTAモデルではなく、Agentic Coding、Cyber、そしてエンタープライズの長文ワークフローに特化したスペシャリスト・モデルである。

モデル	ML4	Llama 4 Maverick	Llama 3.1 405B	GPT-4o
アーキテクチャ	MoE (1.05T Total / 49B Active)	MoE (400B Total / 17B Active)	Dense (405B)	非公表 (Omni)
コンテキスト	1M (実測524k)	1M	128k	128k
ライセンスと 透明性	未公表 (月末公開予定)	Llama 4 Community	Open weights (Model Card有)	Closed (Proprietary)

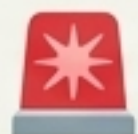
インサイト：ML4は、Llama 4 Maverickが目指す『**MoE・マルチモーダル・ロングコンテキスト路線**』をさらに**大容量化した**対抗馬としての位置づけ。

Model Weights Requirements

Hardware Targets



Active 49BだからといってVRAMが98GBで済むわけではない。
総パラメータの常駐とExpert間の通信帯域がボトルネックになる。



PREVIEW API 利用における重大な法的リスク



オプトアウト不可の学習データ利用条項

- 現在のML4 APIは『Mistral Commercial Terms』の Labs/Preview特則が適用される。
- 入力した顧客データ (Customer Data) および出力 (Outputs) は、Mistralのモデル学習に利用される可能性がある。

Actionable Advice

機密コード、M&A資料、患者情報などをPoCとしてそのままPreview APIに投入することはセキュリティ上避けるべき。イト版のセルフホストを待つこと。

適材（エンタープライズ導入に最適）



• Cybersecurity SOC / インシデント対応（IR）

• リポジトリ規模のコーディング・エージェント

• 法務・財務の大量文書処理

• データ主権を重視する欧州公共機関

不適・オーバースペック（TCOが合わない）



• 単純な社内Chatbot

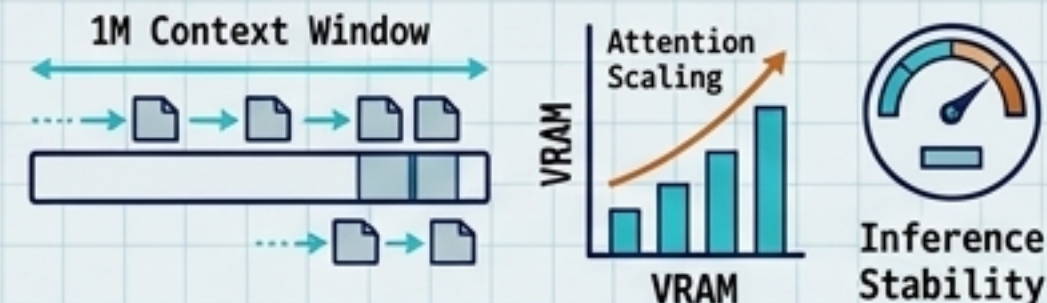
• 小規模なRAG（Retrieval-Augmented Generation）システム

理由：1.05Tのストレージ要件が巨大。短文検索であれば、小型モデル+RAGによるコンテキスト絞り込みの方がレイテンシとコスト面で有利。

ウェイト公開直後の検証チェックリスト (The Verification Playbook)

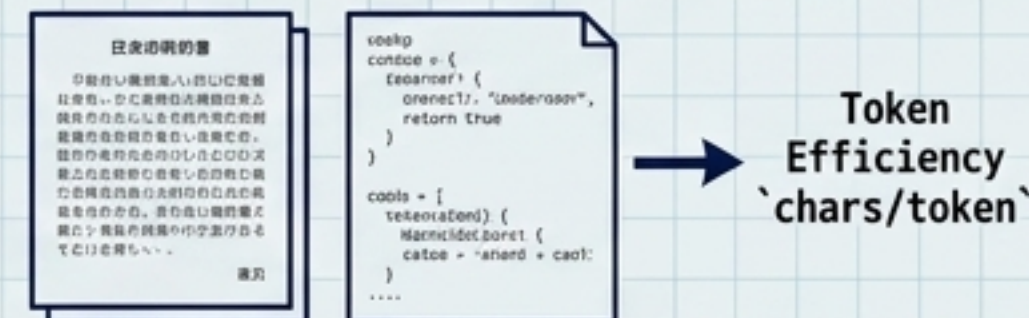
☐ 1Mコンテキストの真正性

Needle-in-a-Haystackのみならず、`1M`到達時の実質的な推論維持力と`KVキャッシュ`の`VRAM消費量` (attention scaling) を測定。



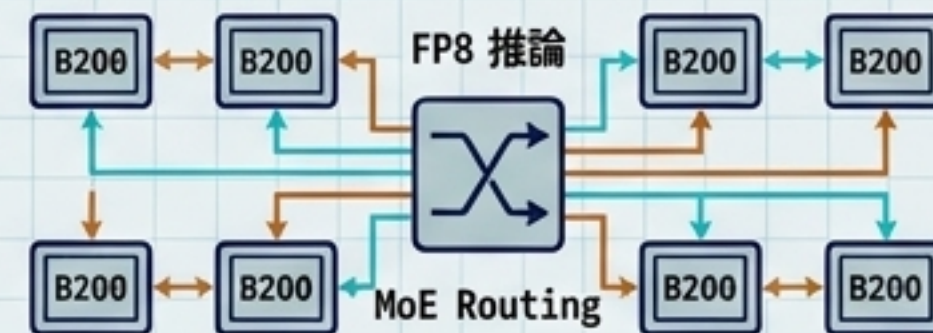
☐ 日本語実効性能

「`160+言語`」の均等性は不明。日本の契約書やコードレビューデータでの`Token効率` (chars/token) を実測。



☐ ハードウェア最適化

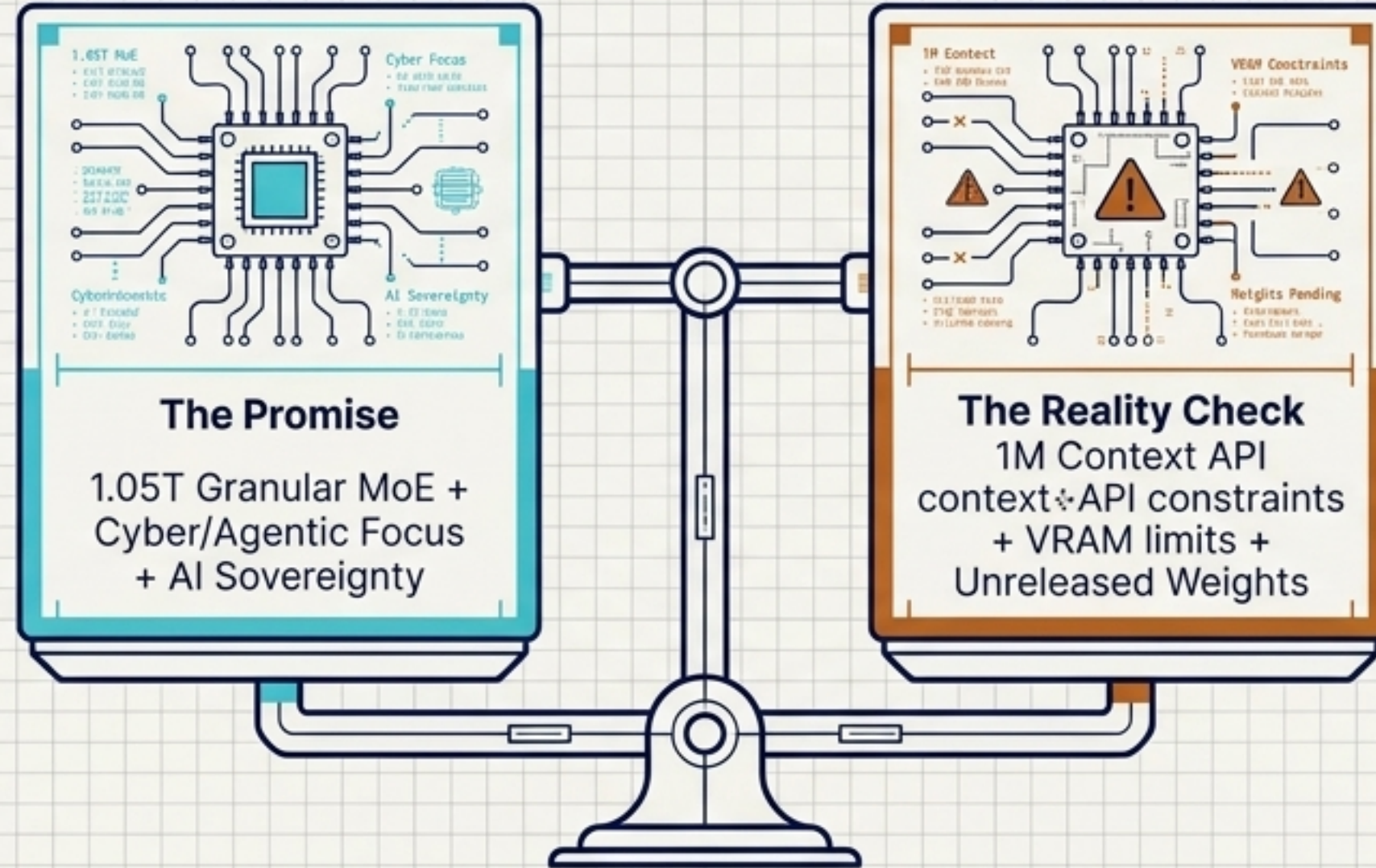
`8×B200`環境での`FP8`/8-bit`推論効率。`MoE`特有のノード間／`GPU間ルーティング通信`のプロファイリング。



☐ ライセンス確認

オープンウェイト化に伴う商用利用ライセンスの制約事項の法務確認。





結論:

欧州発の極めて野心的なフロンティア・モデル。特定のエンタープライズ領域においてGPT-4oを代替し得るポテンシャルを持つが、スペックだけで全面導入を判断すべきではない。真価は月末のウェイト公開と、自社インフラでの物理的な独立検証に委ねられる。