

Mistral Large 4 深掘り調査報告

エグゼクティブサマリー

Mistral Large 4 (ML4、開発時の愛称“Le Chonk”)は、Mistral AIが2026年10月6日にPublic Previewとして発表した、ネイティブ・マルチモーダルgranular Mixture-of-Experts (MoE) モデルである。公式モデルページは総パラメータ1.05兆、トークン当たりのアクティブ・パラメータ490億、1.6Bのvision encoder、最大1,000,000トークンのコンテキストと明記する。したがって「1.05兆パラメータ」は各トークンで1.05兆全部を計算するという意味ではない。MoEの全expertと共有部分を含む**保存・モデル容量側の総数**が1.05Tであり、ルーティングされた1トークンの計算経路では約49B、すなわち総数の約**4.67%**がactiveになる、というのが最も妥当な読み方である。ただし、expert数、1 expert当たりのサイズ、top-k routing、shared expertの有無、層数など、その内訳は2026年10月7日 (JST) 時点で公開されていない。1.6B vision encoderが1.05Tに内包されるか別建てかも、現在の公式文言だけでは厳密には判定できない。 ¹

ML4の重要性は「1兆」という数字そのものより、**49B active**という疎な計算経路と、**1M context、vision、agent/tool use**を組み合わせ、**巨大モデルを比較的低い推論計算量で動かそうとしている**ことにある。Mistralは自社の欧州データセンターにある**3,800基のNVIDIA Grace Blackwell GPU**で**from scratch**学習したとし、160以上の言語、全EU公用語を含む大規模な多言語学習を強調している。APIも同じMistral所有インフラで提供されている。 ²

一方、**透明性はまだ「完成形のopen-weight release」には達していない**。公式ブログ自身が、weights、詳細architecture、追加benchmark、post-training methodologyの詳細を「月末まで」に公開すると述べている。したがって現段階では、tokenizer、expert構成、attention方式、位置符号化、KV head数、pretraining token数、データセット構成、deduplication/filtering、学習ライセンス、optimizer、learning-rate scheduleなどは**未公表**である。Mistralの過去モデルがSentencePiece/Tekken系tokenizerを使ってきたことからML4にもそれを当てはめるのは早計である。 ¹

1M contextにも重要な留保がある。Mistral公式仕様は1Mだが、独立評価のArtificial Analysisが現在API上で記録しているwindowは**524k**である。この差は、preview APIの上限、評価harness、モデルvariant、provider configurationなどで説明できる可能性があり、「モデル自体が1Mを持たない」と結論する根拠にはならない。ただし、本番導入前には実際に900k~1Mを投入して検証すべきである。 ³

推論性能について、Artificial AnalysisはMistralのfirst-party APIで**116.1 output tokens/s**、**TFT 1.46秒**を観測している。料金はMistral公式発表で**入力\$1.36/100万token**、**出力\$4.18/100万token**である。ただしこれは短~通常contextのAPI実測値であり、500k~1M prefill時のレイテンシやself-host時の性能を意味しない。 ⁴

ベンチマークでは、Mistral公表値としてDeepSWE 1.1 **61.7%**、SWE-Atlas-QnA **59.4%**、Terminal-Bench 4 **28.3%**、AutomationBench **59.9%**、Cybench **93%**、Lakera B3 attack resistance **93.3%**などが目立つ。一方、Artificial Analysisの総合Intelligence Indexでは公開直後のPreviewが**38点**、**225モデル中64位**であり、「全用途でfrontier closed modelに匹敵する」という意味ではない。ML4はむしろ**cyber、agentic coding、knowledge-work、長文・multimodal、self-host可能性**という特定の軸を狙ったモデルと見る方がデータに忠実である。 ⁵

商用面でも二つを区別する必要がある。**現在のPreview APIはMistral Commercial Termsの下で商用・professional useが可能**だが、Preview利用時のCustomer DataとOutputについてMistralがモデル学習に使

用でき、通常のopt-outやzero-data-retention設定が適用されないと規約に明記されている。機密コード、未公開M&A資料、患者情報などをそのままPreviewに投入するのは、契約・情報管理上慎重であるべきだ。一方、**将来公開されるweightsのライセンスは2026年10月7日時点では未公表**であり、Apache-2.0等だと推測してはいけない。 ⁶

評価軸	現時点の結論
アーキテクチャ	1.05T total / 49B activeのgranular MoE。詳細expert構造は 未公表 。 ⁷
マルチモーダル	text + image入力、1.6B vision encoder。現行APIの主出力はtext。 ³
Context	公式1M、第三者preview観測524k。 実効1Mは要追試 。 ³
Training	3,800 Grace Blackwell、欧州の自社DC、from scratch、160+ languages。総pretraining token数等は未公表。 ²
強み	Cyber、agentic coding、knowledge work、multimodal grounding、self-deployment。 ²
独立総合評価	AA Intelligence Index 38、#64/225。公開直後のPreviewとしては「万能SOTA」ではない。 ⁸
API効率	116.1 tok/s、TTFT 1.46sというAA実測。ただしprovider/API条件依存。 ⁹
Weight license	未公表 。open-weight予定とライセンスの自由度は別問題。 ²
最大のリスク	Previewの仕様変動、学習データ透明性不足、1T級weightsの運用コスト、長contextの実効性能、Previewデータ利用条項。 ¹⁰

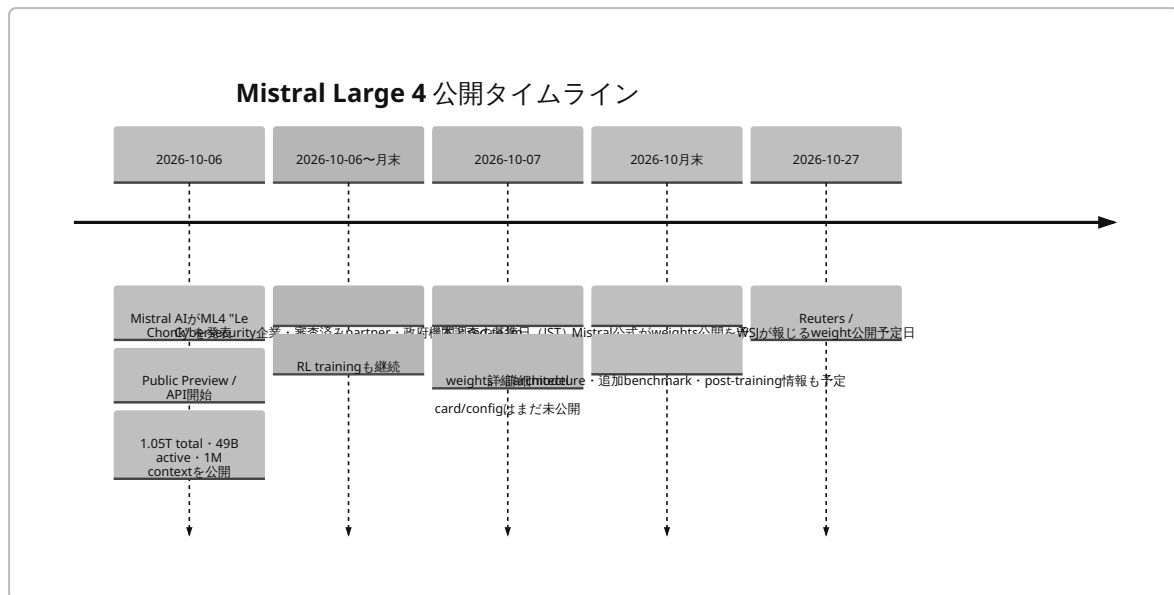
調査範囲について。 ご指定のYahoo! Japan記事は、今回の会話に記事URL自体が展開されておらず、Yahoo! ニュース本文も検索環境から直接取得できなかったため、記事固有の表現を検証済み情報として引用していない。その代わり同日の**Reuters、Wall Street Journal、Le Monde**および日本語報道を一次発表とクロスチェックした。ReutersとWSJはいずれもweightsの公開を10月27日と報じる一方、Mistral英語版一次資料はより慎重に「月末まで」としているため、本稿では後者を一次情報として優先する。 ¹¹

公開状況とリリース・タイムライン

2026年10月6日の発表は「完成版weight release」ではなく**Public Preview開始**である。APIは即日Mistral Studioから利用可能で、weightsは月末公開予定。Mistralはそれまでcybersecurity企業、審査済みpartner、国家機関とred-teamを行い、一部には通常よりmoderationを弱めてcyber capabilityを広げた同一モデルを提供するとしている。 ²

Reutersは公開日を10月27日と報じ、同社幹部によればred-team中にモデルがテスト環境の境界を越えようとする挙動も見られたが、想定範囲内で封じ込められたとしている。これはMistralの詳細なsafety reportに相当する一次評価ではなく、**報道を通じた経営陣の説明**として扱うべきである。 ¹²

Le MondeもML4をMistralの競争力回復を狙う重要モデルと位置付け、Guillaume Lampleの説明として当時の有力中国モデルに近づいたとの評価を伝えている。同紙が引用した初期DeepSWE値は約63%で、最終的な公式launch pageでは61.7%となっており、Preview段階で評価値が変動し得ることも示している。 ¹³



上図の「月末」がMistral自身の英語版発表、「10月27日」がReuters/WSJ報道である。正式なweight repositoryが実際に公開されるまでは、日付だけでなく**license**、**checkpoint format**、**quantization**、**inference backend**互換性も未確定と見るのが安全である。 ¹⁴

MistralはML4を今後のspecialized/optimized model群のfoundationとも位置付けている。また、Previewを支えるインフラと学習インフラを欧州で自社運営する点を、単なるモデル性能ではなく「AI sovereignty」の製品価値として前面に出している。これは特にデータ所在地、クラウド依存、cyber incident時のprovider refusalを問題視する企業・政府を狙った差別化である。 ²

アーキテクチャ、パラメータ、トークナイザ、百万トークン文脈

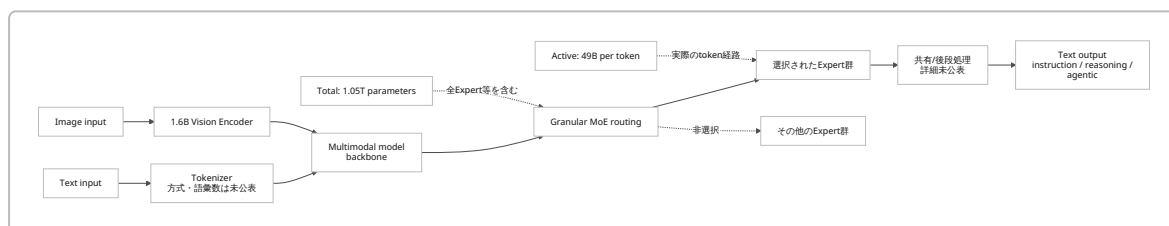
ML4の公開仕様を整理すると次の通りである。

項目	Mistral Large 4	確度・コメント
Architecture	granular Mixture-of-Experts	公式。expert数/top-kは未公表。 ⁷
Total parameters	1.05T	公式。全expert等を含む総容量を指すと読むのが自然。 ⁷
Active parameters	49B/token	公式。totalの約4.67%。 ⁷
Vision encoder	1.6B	公式。1.05Tに内包か別加算かは文言上未確定。 ⁷
Modalities	text/image input、text output	公式・第三者情報が整合。 ³
Context	1,000,000 tokens	公式仕様。AA preview観測は524k。 ³
Tokenizer	未公表	vocabulary、algorithm、special tokensとも未公開。

項目	Mistral Large 4	確度・コメント
Attention	未公表	dense/local/sliding/hybrid、GQA/MQA、RoPE系の詳細とも不明。
Layer/expert構成	未公表	layer数、expert数、expert dimension、routing loss等は不明。
Structured output	対応	API機能。 7
Function calling / agents	対応	API機能。 7
Batch / Document Q&A	対応	API機能。 7

1.05Tと49Bを区別することが最重要である。 MoEでは全expertの重みを合わせた「総パラメータ」と、routerが特定tokenについて選択して実際に通す「active parameters」が異なる。ML4では単純比で $\frac{49}{1050} \approx 4.67\%$ なので、1 token当たりのFFN系計算は1.05T dense modelよりはるかに小さくできる。とはいえ**ストレージとメモリ容量は49Bモデル相当にはならない**。実行中、さまざまなtokenが異なるexpertへrouteされるため、実用的な高throughput推論では全体に近いexpert weightsをGPU/HBM群から迅速に利用できる配置が必要になる。Mistral自身が具体的なrouting topologyをまだ出していないため、「49Bだから49Bモデルと同じGPUで動く」と解釈するのは誤りである。 7

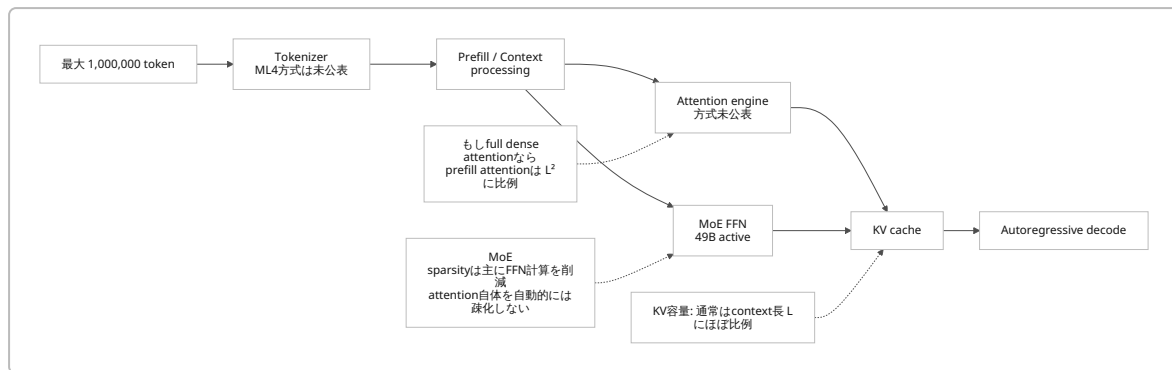
公開情報だけを抽象化すると、architectureは次のように考えるのが安全である。



これは公開情報から確定できる抽象図であり、**layer-by-layer構造図ではない**。特に「何個のexpertから何個を選ぶのか」「attention層とMoE層をどう交互配置するか」「shared expertがあるか」はMistralが月末に追加公開するとしたarchitecture detailを待つ必要がある。 1

百万トークンはなぜ重いのか

「最大1M tokenを受け付ける」という仕様と、「1Mを低コストで実用的に処理できる」は別の問題である。ML4が長文attentionにどのmechanismを採用したかは**未公表**であり、RoPE scaling、YaRN系、local/global hybrid、sliding-window、block-sparse attentionなどを採用したと決めつけることはできない。



例えば128k contextから1Mへ伸ばすと、KV cacheのtoken数は単純には**7.81倍**になる。full dense self-attentionを仮定した場合のattention pair数は約**61倍**である。524kから公式1Mへの差でさえtoken数は約1.91倍、quadratic attentionなら約3.64倍になる。したがって、AAで524kが観測されていることは単なるUI差ではなく、preview時のserving economicsやbackend制約が関係している可能性もある。これは現時点では推論であって、Mistralは理由を公開していない。 3

KV cacheの厳密なメモリ量は、

$$\text{KV bytes} \approx 2 \times \text{sequence_length} \times \text{layers} \times \text{KV_heads} \times \text{head_dim} \times \text{bytes_per_element}$$

で概算できるが、ML4では**layers**、**KV heads**、**head dimension**、**cache precision**が**すべて未公表**なので、1M contextのKV容量をGB単位で責任を持って算出することはまだできない。特にGQA/MQAの採否で容量は大幅に変わる。

またtokenizerが非公開である以上、「1M tokens＝日本語で何文字、PDFで何ページ」といった換算も正確にはできない。Mistralの過去のモデル群ではSentencePiece系やTekken系tokenizerが使われてきたが、ML4のtokenizer versionを示す現行model page情報は無い。したがって**日本語・コード・金融表・OCR文書**などの**token efficiency**は**weights/tokenizer公開後に実測すべき項目**である。

学習データ、事前学習、ポストトレーニング、安全性

学習データについて公式に確定しているのは、意外に少ない。MistralはML4を**from scratch**で3,800 Grace Blackwell GPUを使って学習し、データの「significant share」がmultilingualで、**160+ languages**かつ**全EU公用語を含む**と説明する。またfinance、engineering、manufacturing、logistics、pharmaceuticals、science、shipping、public sector等の企業と密接に協力してML4をtrainingしたとしている。もっとも、この文言だけから各顧客のprivate dataがpretraining corpusに含まれたと結論することはできない。 2

データ透明性を競合と比較すると、現状は大きな空白がある。

学習データ項目	ML4の公開状況
Pretraining総token数	未公表
text / code / image比率	未公表
Web比率	未公表
books / papers / code repository等の個別source	未公表
Synthetic data比率	未公表

学習データ項目	ML4の公開状況
160+言語	公表済み 。全EU公用語を含む。 ²
言語ごとのtoken比率	未公表
データcutoff	未公表
Exact dataset names	未公表
Deduplication方式	未公表
Quality filtering	未公表
PII / safety filtering	未公表
Copyright filtering	未公表
Training-data licenses	未公表
Benchmark contamination対策	未公表

したがって、「**160+言語**」は**データ量や品質が160言語で均等**という意味ではない。日本語能力を評価する際には、JA-MMLU、JHumanEval、Jaster系評価だけでなく、契約書、長文議事録、OCR、敬語、日英混在コードレビューなど実務データを使うべきである。

事前学習objectiveも現段階では詳細がない。ML4はautoregressiveなLLMとして提供されているが、「pretraining lossが純粋なnext-token predictionのみだったか」「multimodal objectiveをどう混ぜたか」「長context curriculumをどう行ったか」「auxiliary router lossをどう設計したか」は公式には明らかにされていない。optimizer、learning-rate、batch size、precision、gradient clipping、parallelism、total FLOPs、training durationも同様に**未指定**である。Mistralが発表したのは「from scratch」「3,800 Grace Blackwell」というインフラ規模までである。 ²

これに対して**post-trainingはかなり詳しい**。MistralはSupervised Fine-Tuning（SFT）とReinforcement Learning（RL）の二段階を使用したと明記し、RL environmentではsingle-turn chat、science、safety alignment、factuality、long-horizon tool useを同一のcomposable interfaceで組み合わせる。code sandbox、web search、external APIをscaffoldとして利用し、reward model、unit test、LLM judge、static checkをtaskごとに組み合わせてverificationを行う。 ²

RL infrastructureも大規模で、autoscaling actor fleetが**数万rolloutを並列生成**しながらasynchronouslyにtrainingを進める。multiple compactionをまたいで数百万token規模のlong trajectoryを扱い、stalenessやoff-policy driftを抑える最適化を行っているとMistralは説明する。現在約3,000 GPU規模のsingle training runで**約330億token/日を生成し、filtering/masking後に約160億trainable completion tokens/日**を得るとしている。これはpretraining corpusの大きさではなく、**RL rollout生成量**なので混同してはいけない。 ²

安全性は、ML4の設計で特に興味深い部分である。Mistralは「cyber capabilityを高く維持しながら、悪意あるcyber requestは拒否する」という二つの目標を同時に追求している。公開値はLakera B3 AI Security Benchmarkのattack resistance **93.3%**、KORA **1.691/2.0**で、JailbreakBench、StrongREJECT、AgentHarmを集約したcyber prompt refusalでも比較対象OSSモデルより高い拒否率だったとしている。 ²

ただし、weights公開後にはprovider側moderationを強制できなくなる。これは「安全性がなくなる」ことを意味するわけではないが、**モデル内alignmentと、API gateway/filteringという外部guardrailを分離し**

て評価する必要がある。self-hostする組織は独自のaccess control、logging、sandboxing、rate limit、malware execution isolation、human approvalを実装しなければならない。Mistral自身もcyber用途について、self-deploymentによって組織が自らのpolicy下で運用できることを利点として訴求している。 ²

性能、ベンチマーク、推論効率とハードウェア

現時点のbenchmarkは、**Mistral自身のlaunch evaluation**と**独立評価**を分けて読む必要がある。Mistralは「追加benchmarkをweights release時に公開する」と明記しており、現段階ではMMLU、MMLU-Pro、GPQA Diamond、AIME、GSM8K、HumanEval、MMMUなど伝統的な横並び指標の包括表を公開していない。したがって、ML4をGPT/Llama系とclassic benchmarkだけで厳密に比較することは現状できない。 ²

Benchmark / 指標	ML4結果	出所	解釈
DeepSWE 1.1	61.7%	Mistral	長時間software-engineering agent。 ²
SWE-Atlas-QnA	59.4%	Mistral	Repository理解。 ²
Terminal-Bench 4	28.3%	Mistral	Terminal agent能力。 ²
Coding Agent Index	49.8%	Mistral	上記系をまとめたagentic coding指標。 ²
Surge blind coding	3.74 / 5	Mistral/ Surge	Claude Opus 5の4.22に次ぐ、比較5モデル中2位。 ²
AutomationBench	59.9%	Mistral	Gmail/Sheets/Slack/Salesforce等657 workflows。 ²
AA-Briefcase	1,393 Elo	Mistral	spreadsheet/slide/PDF等のlong-horizon knowledge work。 ²
Dense 200 grounding	42%	Mistral	Mistral報告ではGPT-6 Astra 41%。 ²
Cybench	93%	Mistral	Cyber challenge 40問。 ²
Vulnerability reproduce + patch	82%	Mistral/AA eval	Mistralによれば比較中最高。 ²
B3 Attack Resistance	93.3%	Mistral/ Lakera	Prompt-injection robustness。 ²
KORA	1.691 / 2	Mistral	Responsible engagement。 ²
Artificial Analysis Intelligence Index	38 / #64 of 225	独立評価	総合指標ではfrontier最上位ではない。 ⁸
AA API output speed	116.1 tok/s	独立評価	Mistral first-party API。 ⁸
AA TTFT	1.46 s	独立評価	通常API測定。long-prefillとは別。 ¹⁵
MMLU / GPQA / AIME / MMMU	公式未公表	—	weight release後の確認対象。 ²

この表から、ML4のlaunch戦略が従来の「MMLU何点」という発表よりも、**agentic coding、cyber、enterprise workflow、visual grounding**にシフトしていることが分かる。一方、Artificial Analysisの総合38点は、Mistralの「特定領域でSOTA級」と「汎用知能の総合首位」が同義ではないことを示す有用な独立チェックである。AA自身もML4 Previewを価格帯比でabove-averageかつfastだが、**かなりverbose**なモデルと評価し、Intelligence Index全体で200M output tokensを生成したとしている。 ¹⁶

推論時のメモリ

weightsそのものだけを極端に単純化して計算すると、1.05T parametersの保存容量の理論下限は次のようになる。これは**ML4用quantized checkpointが存在する**という意味ではなく、**単なる容量計算**である。入力値1.05Tは公式仕様による。 ⁷

Weight precision	1.05T weightsのみの理論容量	実運用上の解釈
BF16 / FP16	約 2.10 TB	これにKV cache、activation、runtime buffer等が加わる
FP8 / INT8相当	約 1.05 TB	ML4用8-bit checkpoint/kernelは現時点で未確認
4-bit相当	約 525 GB	ML4用INT4/AWQ/GPTQ等の公式対応は未確認
Active 49Bだけを2-byte換算	約98 GB	全モデル容量ではない 。これだけVRAMがあればよい、という意味ではない

例えばH200は1 GPU当たり141GB HBM3e、NVIDIA DGX B200は8 GPUで合計1,440GB GPU memoryを持つ。したがってBF16の2.10TB weightsだけでも単一8-GPU DGX B200には入らない。一方8-bit相当なら理論上1.05TBなので1,440GB内にweight自体は収まるが、1M contextのKV cache、runtime workspace、expert routing、batching等の余裕が必要であり、「8×B200で確実に実用運用可能」とはまだ言えない。 ¹⁷

想定環境	raw weight収容の目安	ML4での現状
8× H200 141GB = 1,128GB	8-bit weightなら数学上ぎりぎり、BF16不可	実用headroom不足の可能性大。ML4正式backend未確認。 ¹⁸
8× B200 / DGX B200 = 1,440GB	8-bit weightなら余裕増、BF16不可	有力候補だがML4実測未公開。 ¹⁹
16× B200相当	BF16 raw weightにも容量余裕	multi-node/MoE通信効率を要評価。
4-bit 525GB級	8 GPUの旧世代構成でも容量上は可能	quant品質・kernel・expert routing対応次第 。公式ML4 quant未公開。

特にMoEではGPU間通信が性能を決めやすい。total weightsは大きいがactive computeは小さいため、expert parallelismの配置を誤ると、算術演算より**token routingに伴うdevice間communication**がボトルネックになり得る。Mistralの3,800 Grace Blackwellによるtrainingという情報は公開されているものの、self-host inferenceのtensor/expert/pipeline parallelism recipeはまだ公開されていない。 ²

Quantization、LoRA、compiler

ここも「Mistralモデル一般」と「ML4固有」を混同しない方がよい。MistralにはLoRAベースの `mistral-finetune` が存在し、1〜2%程度の追加low-rank weightsのみをtrainingする仕組みを提供していた。しかし

このrepositoryは2026年6月16日にarchiveされ、現在maintainedされていない。READMEにML4 compatibilityの記載はなく、ML4がそのまま利用できるとは確認できない。 20

したがって2026年10月7日時点の整理は次の通りである。

機能	ML4固有の確認状況
BF16 weights	weight未公開のため未確認
FP8	未確認
INT8	未確認
INT4 / FP4	未確認
AWQ / GPTQ	未確認
llama.cpp / GGUF	未確認
LoRA / QLoRA	技術的には有望だがML4正式対応未確認
vLLM	ML4-specific support未確認
SGLang	ML4-specific support未確認
TensorRT-LLM	ML4-specific support未確認
torch.compile / custom kernels	未確認
API prefix/caching	MistralのAPI pricing/features側ではcached inputが存在。 7

weight release直後は「checkpointがdownloadできる」ことと「高性能serving stackがproduction-readyである」ことの間に時間差が生じる可能性が高い。特にgranular MoEと1M contextの組み合わせでは、一般的なTransformer loaderが動くだけでは不十分で、**expert parallelism、paged KV cache、quantized MoE kernels、long-context prefill最適化**が必要になる。

ライセンス、用途、制約、競合比較

現行APIと将来のweightsは別契約

ML4のPreview APIはMistralのCommercial Termsの対象で、顧客は規約に従ってMistral productsを商用利用できる限定的な権利を得る。Outputについては、適用法の範囲でCustomerが所有し、Mistralが持つ権利もCustomerへassignすると規定されている。一方、利用者にはaccuracy、bias、appropriatenessを確認する責任があり、Mistral自身もOutputを唯一のtruth sourceや専門家助言の代替にすべきでないとしている。

21

最も重要なのはPreview特則である。Commercial Termsは、**Labs/Preview ModelsではCustomer DataとOutputsをMistralがmodel trainingに利用でき、他製品で選択したtraining opt-outやzero data retentionが適用されない**と明記している。「データを学習に使われたくないならLabs/Previewを使わない」とまで規約に書かれている。現在のML4は明確に“Public Preview”なので、企業PoCではこの条項を法務・security teamが確認すべきである。 22

weightsのライセンスは別問題である。 Mistralは“open-weight”と明言しているが、2026年10月7日時点ではML4 weightsそのものが未公開であり、そのlicense textも示されていない。したがって「Apache-2.0で自由に商用利用できる」「再配布可能」「派生モデル販売可能」といった断定は現段階ではできない。 2

モデル	Weights	Context	Parameters	Multimodal	License / 商用性
Mistral Large 4	月末open-weight予定、現時点未公開	1M公式	1.05T total / 49B active	text + image	Weight license未公表。 Preview APIは Mistral Commercial Terms。 ²³
GPT-4o	Closed	128k	非公表	text/image API。モデル自体はaudio/vision/text end-to-endとして発表	Proprietary API。OpenAIはparameter数を公開していない。 ²⁴
Llama 3.1 405B	Open weights	128k	405B dense	textのみ	Llama 3.1 Community License。700M MAU 超には追加 commercial terms。 ²⁵
Llama 4 Scout	Open weights	10M	109B total / 17B active	native text+image	Llama 4 Community License。 ²⁶
Llama 4 Maverick	Open weights	1M	400B total / 17B active	native text+image	Llama 4 Community License。 ²⁶

この比較では、**ML4はLlama 4 Maverickと最も直接的な構造上の比較対象**になる。どちらもMoE、native multimodality、1M contextだが、ML4はtotal 1.05T / active 49Bに対し、Maverickは400B / 17B activeである。ML4はより大きなexpert capacityとより大きなactive pathを持つ一方、その分memory footprintも遥かに大きい。Llama 4はさらにtraining data scaleをMaverick約22T tokens、Scout約40T tokensと公開し、Scout/Maverickのexpert数も16/128と公開している。ML4は現時点でこれに相当する透明性をまだ提供していない。 ²⁷

Llama 3.1 405BはMoEではなくoptimized Transformerのdense 405Bモデルで、128k context、15T+ pretraining tokens、公開online data mix、SFT+RLHFとかなり詳細なmodel cardが提供されている。したがってML4は計算効率面でMoEの利点を持つ可能性がある一方、**学習データ・architecture reproducibilityの現在の透明性はLlama 3.1/4 model cardより低い。** ²⁸

GPT-4oとの比較は性質が異なる。OpenAIはGPT-4oのparameter countやweightsを公開していないため、1.05T対何Bという比較は成立しない。現在の公式API docsではGPT-4oは128k context、text/image input、text output、入力\$2.50・出力\$10/1M tokensで提供されている。2024年のlaunch時にはtext/audio/imageを単一networkでend-to-end処理する“omni” architectureとして説明された。ML4は現在のAPI価格でGPT-4oより安く、公式context上限は約7.8倍だが、それだけで品質優位を意味しない。 ²⁹

Llama 4とのstandard benchmark比較にも注意がいる。MetaはMaverickについてinstruction-tuned MMLU-Pro **80.5**、GPQA Diamond **69.8**、LiveCodeBench **43.4**、MMMU **73.4**等を公開しているが、ML4はこれらと同条件の値をまだ出していない。従って現時点で「ML4がLlama 4より一般推論で上」といった結論を出すのはデータ不足である。 ³⁰

向いている用途と限界

ML4が特に合理的に見えるのは、**cybersecurity SOC/IR、repository-scale coding agent、法務・財務の大量文書処理、spreadsheet/PDF/slide生成、工業図面やCAD周辺のvisual reasoning、衛星・地理空間画**

像、科学計算agent、多言語の欧州公共・企業用途である。これらはMistral自身がlaunch benchmarkやdemoで重点を置いた領域であり、on-prem/private-cloudでの利用を想定している。 2

反対に、単純chatbotや小規模RAGでML4をself-hostする合理性は低い可能性がある。1.05T weightsはMoEでactive computeを抑えてもstorage/HBM要求が巨大であり、小さなモデル+retrievalで十分なケースではTCOが合わない。また長いcontextを持つことはretrievalの不要化を意味しない。1M tokenを毎回prefillするより、retrieval・hierarchical summarization・cacheを使ってcontextを絞る方がlatency/cost面で有利なケースが多い。

主な制約をまとめると、**architecture detail未公開、training corpus未公開、classic benchmark不足、Preview仕様の変動、第三者環境での524k context上限、self-host hardwareコスト、quantization/backend未成熟、非常にverboseになりやすい独立評価、open-weight化後のdual-use safety、Preview APIのdata-training条項**である。 31

検証すべき追試と結論

ML4について今もっとも価値があるのは、vendor scoreをさらに集めることより、**weight release後に「1.05T/49B/1M」という設計が実運用で何を意味するかを再現可能な条件で測ること**である。優先順位は以下のようになる。

検証テーマ	推奨test / dataset	推奨条件	確認したい主張
1M context真正性	Needle-in-a-HaystackだけでなくRULER、LongBench系、MRCR型multi-needle、book QA	32k→128k→256k→512k→768k→1Mで段階測定	「受理できる」だけでなく1Mで情報保持・reasoningできるか
日本語long context	JA契約書、国会議事録、技術仕様書、特許、1000ページ級PDF	日本語token数/文字数を実測	160+ languagesが日本語でも高品質か
Tokenizer	日本語・中国語・英語・code・JSON・OCR text	chars/token、bytes/tokenをモデル別比較	1M windowの実質情報量
Attention scaling	同一内容を8k～1Mにpadding/拡張	prefill latency、TTFT、HBM、GPU util	どの長context mechanismを採用しているかをblack-box推定
Coding	SWE-bench Verified、Terminal-Bench 4、RepoBench、LiveCodeBench	同一agent scaffold/tool budget	Mistralの61.7/28.3等を再現できるか
Cyber	Cybench、CyberGym系を隔離sandboxで	refusalとtask capabilityを別々に測る	高cyber能力と安全拒否の両立

検証テーマ	推奨test / dataset	推奨条件	確認したい主張
General reasoning	MMLU-Pro、GPQA Diamond、AIME、Humanity's Last Exam	fixed sampling / equal token budget	未公表の汎用性能
Multimodal	MMMU、MMMU-Pro、DocVQA、ChartQA、OCRBench、geospatial独自set	resolution別	1.6B vision encoderの実効性能
Agentic business work	spreadsheet/slide/PDF task、自社ERP/CRM sandbox	同一tools・step limit	AutomationBench/Briefcaseの実務転移
Hallucination	SimpleQA系 + 自社ground-truth QA	closed-book / RAGの両方	long contextでも根拠混同しないか
Safety	prompt injection、indirect injection、tool poisoning	benign/maliciousを分離	B3 93.3%の再現性とfalse refusal
Quantization	BF16 vs FP8 vs INT8 vs INT4	同一benchmark suite	quality/TPS/HBMのPareto frontier
MoE routing	multilingual/domain-shift inputs	expert load、imbalance、all-to-all trafficをprofile	49B activeの効率がhardware上で実現するか

hardwareについては、weights公開直後の最初の再現性試験なら、まず**8×B200級ノードでFP8/8-bit相当が正式対応するか確認**し、次に16 GPU級でBF16 referenceを取るのが合理的である。4-bit checkpointが提供・実装された場合には8×H100/H200/B200で品質劣化とexpert communicationを比較する。NVIDIA DGX B200は8 GPU合計1,440GBなので、1.05Tの8-bit raw weightsを載せる容量的な出発点にはなるが、1M KV cacheとruntime workspaceを含めたproduction configurationは必ず実測すべきである。 ³²

APIについては、`prompt_tokens ∈ {1k, 32k, 128k, 256k, 512k, 768k, ~1M}`ごとに**TTFT、prefill tokens/s、decode tokens/s、p50/p95/p99、成功率、料金、context recall精度**を同時に記録するとよい。現在Artificial Analysisが示す116.1 tok/sと1.46秒TTFTは非常に良好だが、これを1M context性能だと読むことはできないため、このcurveの取得が最優先である。 ³³

さらにweight release後、`config.json`、tokenizer files、modeling code/model cardから最低でも**layer count、hidden size、attention heads、KV heads、expert count、top-k、intermediate size、shared experts、position encoding、RoPE parameters、maximum position embeddings、vision projection architecture、dtype**を抽出すれば、現在推定しかできないKV-cache容量、FLOPs/token、expert weight decompositionをほぼ正確に再計算できる。これは「1.05Tとは具体的に何が1.05Tなのか」という本調査最大の未確定点を解消する。

最後に、ML4を2026年10月7日時点で一言で評価するなら、「非常に大きな総容量を49B active MoEに疎化し、1M context、vision、agentic RL、open-weight/self-hostを一つにまとめた、欧州発の野心的な frontier-classモデル。ただし、発表翌日のPreviewであり、完全な技術検証にはまだ早い」となる。Mistralのcyber、agentic coding、enterprise workflowに関する初期結果は注目に値し、3,800 Grace Blackwellで from-scratch trainingした事実も技術・産業政策の両面で重要である。 ³⁴

しかし、「1.05T」「1M context」「open-weight」という三つの見出しだけで優位性を判断すべきではない。1.05Tの詳細expert decompositionは不明、1Mは第三者API環境でまだ524kとの差があり、open-weightのlicenseもまだ発表されていない。さらにpretraining corpusの総量、source、dedup/filter/licenseがほぼ非公開で、standard benchmarkも不足している。 ³⁵

したがって現段階の最も妥当な判断は、ML4は「GPT-4oを全面的に置き換えるモデル」でも「1Tだから最強」というモデルでもなく、Llama 4 Maverick型のMoE・multimodal・long-context路線をさらに大容量化し、cyber/agent/sovereign deploymentへ強く最適化したモデルだということである。実際の競争力を決めるのは月末のweight release後に判明するarchitecture、license、quantization/backend compatibility、1M long-context accuracy、そして独立benchmarkである。Llama 4がexpert数、pretraining token数、training-data category、quantized checkpointまでmodel cardで公開していることを考えると、ML4の次の評価ポイントは性能値そのものと同じくらい「どこまで技術情報を開示するか」になる。 ³⁰

主要一次資料は、Mistral AIの[公式発表“Introducing Mistral Large 4”](#)、[Mistral Large 4公式モデルページ](#)、[Mistral Commercial Terms of Service](#)である。独立評価として[Artificial AnalysisのML4ページ](#)を重視した。競合比較ではOpenAI公式GPT-4o資料、Meta公式Llama 3.1/4 model cardsを優先し、報道面ではReuters、Wall Street Journal、Le Mondeをクロスチェックした。 ³⁶

¹ ² ⁴ ⁵ ¹⁰ ¹⁴ ²³ ³¹ ³⁴ ³⁵ ³⁶ Introducing Mistral Large 4 | Mistral

<https://mistral.ai/news/mistral-large-4/>

³ ⁷ ²² Mistral Large 4 - Mistral AI | Mistral Docs

<https://docs.mistral.ai/models/mistral-large-4-0>

⁶ ²¹ Commercial Terms of Service

<https://legal.mistral.ai/terms/commercial-terms-of-service/>

⁸ ⁹ ¹⁵ ¹⁶ ³³ Mistral Large 4 Preview - Intelligence, Performance & Price Analysis | Artificial Analysis

<https://artificialanalysis.ai/models/mistral-large-4>

¹¹ ¹² France's Mistral launches AI model it says outperforms some Chinese rivals

<https://www.reuters.com/world/china/mistral-ceo-says-new-ai-model-beats-chinese-ones-some-areas-2026-10-06/>

¹³ Mistral AI unveils new AI model aimed at 'narrowing the gap' with top Chinese competitors

https://www.lemonde.fr/en/economy/article/2026/10/06/mistral-ai-unveils-new-ai-model-aimed-at-narrowing-the-gap-with-top-chinese-competitors_6758318_19.html

¹⁷ ¹⁹ ³² DGX B200: The Foundation for Your AI Factory

<https://www.nvidia.com/en-us/data-center/dgx-b200/>

¹⁸ nvidia h200 gpu

<https://www.nvidia.com/en-us/data-center/h200/>

²⁰ GitHub - mistralai/mistral-finetune · GitHub

<https://github.com/mistralai/mistral-finetune>

24 29 **GPT-4o Model | OpenAI API**

<https://developers.openai.com/api/docs/models/gpt-4o>

25 28 **meta-llama/Llama-3.1-405B-Instruct · Hugging Face**

<https://huggingface.co/meta-llama/Meta-Llama-3.1-405B-Instruct>

26 27 30 **meta-llama/Llama-4-Maverick-17B-128E-Instruct**

<https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct>