

Mistral 3: 分散型インテリジェンスの夜明けとオープンウェイトAIの新たな覇権に関する包括的技術調査報告書

Gemini 3 pro

1. 序論: 2025年のAIランドスケープにおける転換点

2025年12月2日、フランスに拠点を置くAIスタートアップMistral AIは、同社の技術的系譜における最も重要なマイルストーンとなる「Mistral 3」モデルファミリーを発表した¹。この発表は、単なる大規模言語モデル(LLM)のバージョンアップという枠組みを超え、クラウド上のハイパフォーマンス・コンピューティング(HPC)からエッジデバイスに至るまでをシームレスに接続する「分散型インテリジェンス(Distributed Intelligence)」という新たなパラダイムの到来を告げるものであった¹。

過去数年間、生成AIの開発競争は「規模の追求」に主眼が置かれてきた。パラメータ数を数兆単位に拡大し、計算資源を無尽蔵に投入することで性能向上を図る「スケーリング則」の信奉が支配的であった。しかし、Mistral AIが提示した「Mistral 3」は、この潮流に対するアンチテーゼとも言える「効率的な規模(Efficient Scale)」と「実用的な知能」への回帰を体現している¹。

本報告書は、Mistral 3ファミリーの中核を成すフラッグシップモデル「Mistral Large 3」およびエッジ向け高密度モデル群「Ministral 3」について、その技術的アーキテクチャ、ハードウェア最適化、ベンチマーク性能、そして産業界への影響を包括的に分析するものである。特に、NVIDIAとの戦略的協業によって実現されたハードウェアとソフトウェアの共進化(Co-design)、Apache 2.0ライセンスによるオープンソース戦略がもたらすエコシステムの変化、そしてLlama 4やGPT-4oといった競合モデルとの比較優位性について、提供された資料に基づき徹底的に詳述する。

2. Mistral Large 3: フロンティアクラスMoEのアーキテクチャ革新

Mistral 3ファミリーの頂点に位置する「Mistral Large 3」は、オープンソースモデルとして初めて、クローズドな最先端モデル(フロンティアモデル)と完全に対等な性能を持つことを目指して設計された。その核心には、計算効率とモデル容量のトレードオフを解消する高度なアーキテクチャが存在する。

2.1 スパースMixture-of-Experts (MoE) の進化

Mistral Large 3の最大の特徴は、総パラメータ数6,750億(675B)という巨大な知識容量を持ちながら、推論時に活性化されるパラメータ(Active Parameters)をわずか410億(41B)に抑えている点に

ある¹。これは、全パラメータの約6%に過ぎない。この劇的な効率化を実現しているのが、スパースなMixture-of-Experts (MoE) アーキテクチャである¹。

従来のデンス (Dense) モデルでは、入力されるすべてのトークンに対してニューラルネットワーク全体が計算に参加するため、モデルの大規模化に伴い推論コストが線形に増大するという課題があった。対してMistral Large 3のMoE構造は、モデルを多数の専門家(エキスパート)サブネットワークに分割し、ルーター (Router) と呼ばれるゲート機能が入力トークンの内容に応じて最適な専門家のみを選択的に活性化させる¹。

2.1.1 粒度の高いMoE (Granular MoE) と専門家並列化

Mistral Large 3では、「粒度の高い (Granular)」MoEアプローチが採用されている¹。これは、専門家ネットワークをより細かく、多数に分割することで、知識の専門化を促進し、ルーターによる選択精度を向上させる手法である。この粒度の高さは、特定のタスク(例えばコーディング、創造的執筆、多言語翻訳など)に対して、より特化したニューロン群を動員することを可能にし、計算リソースの無駄を排除しながら回答の質を高めることに寄与している³。

さらに、このアーキテクチャはNVIDIAのGPU間相互接続技術であるNVLinkを活用した「広帯域エキスパート並列化 (Wide Expert Parallelism)」に最適化されている¹。通常、MoEモデルでは専門家が異なるGPUメモリ上に分散して配置されるため、トークンのルーティングに伴う通信オーバーヘッドがボトルネックとなり得る。しかし、Mistral Large 3は、NVIDIAのコヒーレントメモリアクセス機能を前提に設計されており、GPUクラスター全体を単一の巨大なメモリプールとして扱うことで、この通信遅延を極小化している¹。

技術仕様	Mistral Large 3	解説・含意
アーキテクチャ	Granular Sparse MoE	高い専門性と計算効率の両立を実現する設計思想。
総パラメータ	675B (6,750億)	モデルが保持する総知識量。GPT-4クラスに匹敵する規模。
アクティブパラメータ	41B (410億)	推論1トークンあたりの計算量。運用コストは中規模モデル並み。
コンテキスト長	256,000トークン (256K)	書籍数冊分、大規模コードベースを一括処理可能 ¹ 。

データ型	NVFP4 / FP8 / BF16	最新GPU(Blackwell)の性能を引き出す低精度演算に対応 ¹ 。
------	--------------------	---

1

2.2 マルチモーダルネイティブとコンテキスト処理能力

Mistral Large 3は、テキスト処理能力に加え、視覚情報(Vision)をネイティブに理解するマルチモーダル能力を備えている¹。これは、外部の画像エンコーダーを後付けで統合する従来の手法とは異なり、事前学習段階からテキストと画像を同一の埋め込み空間で処理するように訓練されていることを示唆している。これにより、画像を含む複雑なドキュメントの解析や、視覚情報を伴う推論タスクにおいて高い整合性を発揮する⁸。

また、256,000トークン(256K)というコンテキストウィンドウは、企業ユースケースにおいて決定的な意味を持つ¹。RAG(Retrieval-Augmented Generation)システムにおいて、従来はドキュメントを細切れのチャンクに分割して検索する必要があったが、Mistral Large 3の長大なコンテキスト容量により、関連するドキュメント全体や過去の長い対話履歴をそのままプロンプトに入力することが可能となる¹⁰。これにより、情報の欠落(Lost in the Middle現象)を防ぎ、より正確で文脈に沿った回答生成が可能となる。

3. NVIDIAとの戦略的協業: ハードウェアとソフトウェアの共進化

Mistral 3の性能を語る上で不可欠な要素が、NVIDIAとの深いレベルでの技術提携である。このパートナーシップは、単にGPUを購入するという関係を超え、次世代ハードウェアアーキテクチャとAIモデルの設計を相互に最適化する「Co-design(共設計)」の領域に達している¹。

3.1 GB200 NVL72とBlackwellアーキテクチャへの最適化

Mistral Large 3は、NVIDIAの最新プラットフォームである「GB200 NVL72」システム上でその真価を発揮するように調整されている¹。GB200 NVL72は、72基のBlackwell GPUをNVLink Switchで相互接続し、毎秒130テラバイト(TB/s)という驚異的な帯域幅で結合したラックスケールのシステムである。

Mistral AIはこのハードウェア特性を前提に、MoEの専門家配置とルーティングアルゴリズムを最適化した。具体的には、NVLinkのコヒーレントメモリドメイン(Coherent Memory Domain)を活用することで、72基のGPUメモリ全体をあたかも単一の巨大なGPUメモリであるかのように扱い、どのGPU上の専門家に対しても低遅延でアクセスすることを可能にした¹。この結果、前世代のH200システムと比較して、推論パフォーマンスは10倍に向上している¹。これはハードウェアの進化分(通常2~3倍)を大きく上回る数字であり、ソフトウェア側のアーキテクチャ最適化が大きく寄与していることを示して

いる。

3.2 NVFP4量子化:精度と速度の限界突破

AI推論の効率化における最大の課題の一つは、計算精度の低下(量子化)に伴うモデルの劣化を防ぐことである。Mistral Large 3は、NVIDIA Blackwellアーキテクチャで初めて導入された「NVFP4(4ビット浮動小数点)」形式に対応したチェックポイントを提供している¹。

従来のFP16(16ビット)やFP8(8ビット)と比較して、NVFP4はメモリ使用量をさらに半減させ、計算スループットを倍増させる。一般に4ビットまで精度を落とすとモデルの回答品質が著しく低下するが、Mistral AIとNVIDIAは、モデルの重みの分布特性に合わせた高度な量子化アルゴリズムと、Blackwell GPUのTensor Coreによるハードウェア支援を組み合わせることで、「精度を維持したまま(accuracy-preserving)」の4ビット推論を実現した¹。これにより、675Bという巨大なモデルであっても、より少ないGPU数で、かつ高速に運用することが可能となり、トークンあたりの生成コストを劇的に削減している。

3.3 NVIDIA Dynamoと分離型推論(Disaggregated Inference)

大規模な推論システムにおけるもう一つの革新が、「NVIDIA Dynamo」フレームワークを活用した「分離型推論(Disaggregated Inference)」の採用である¹。

LLMの推論プロセスは、プロンプトを一括処理する「プリフィル(Prefill)」フェーズと、1トークンずつ生成する「デコード(Decode)」フェーズに分かれる。プリフィルは計算集約型(Compute-bound)であり、デコードはメモリ帯域集約型(Memory-bound)であるため、これらを同一のGPUで行うとリソースの競合が発生し、効率が低下する。

NVIDIA Dynamoは、この2つのフェーズを異なるGPUリソースに分離し、ネットワーク経由でパイプライン処理を行うことを可能にする。Mistral Large 3はこの仕組みにネイティブ対応しており、例えばプリフィル専用のGPU群とデコード専用のGPU群を動的に割り当てることで、サーバーの稼働率を最大化し、ユーザーに対する応答遅延(レイテンシ)を最小化している⁷。

4. Mistral 3: エッジAIにおける「知能」の再定義

Mistral AIのビジョンである「分散型インテリジェンス」において、クラウド上のLargeモデルと対をなす重要なピースが、エッジデバイス向けの「Mistral 3」シリーズである。これらは、限られた計算資源の中で最大限の知能を実現することを目指した高密度(Dense)モデル群である¹。

4.1 モデルバリエーションとスペック詳細

Mistral 3は、用途とハードウェア制約に応じて3つのサイズで展開される⁵。これらはMoEではなくDenseアーキテクチャを採用しており、これはメモリアクセスのパターンが予測しやすく、エッジデバイスのNPUやコンシューマー向けGPUでの最適化が容易であるためと考えられる。

モデル名	パラメータ	特徴・ターゲットデバイス	推奨用途
Ministral 3 3B	30億	スマートフォン、IoT、ドローン	常時起動のアシスタント、リアルタイム翻訳、コマンド処理 ¹¹ 。
Ministral 3 8B	80億	ラップトップ、組み込みPC	オフライン文書作成、要約、基本的なコーディング支援 ¹² 。
Ministral 3 14B	140億	ワークステーション、ハイエンドPC	複雑な推論、マルチモーダル分析、ローカルRAG ¹⁰ 。

4

特に**Ministral 3 14B**は、以前の「Mistral Nemo 12B」の後継として位置づけられ、同サイズ帯において「クラス最高(Best-in-class)」の性能とコストパフォーマンスを誇る¹⁰。パラメータ数を14Bまで微増させたことで、推論能力と知識の定着率が大幅に向上しており、単なる軽量モデルではなく、実務に耐えうる「コンパクトなパワーハウス」として設計されている¹⁰。

4.2 「Reasoning (推論)」バリエーションとSystem 2思考

Ministral 3シリーズの最大の革新点は、各サイズにおいて「Base」「Instruct」に加え、**「Reasoning」**バリエーションが提供されていることである⁴。

このReasoningモデルは、複雑な問題を解く際に、即座に回答を出力するのではなく、内部的な思考プロセス(Chain of Thought)を経てから結論を導き出すように訓練されている¹⁵。これは認知科学における「システム2(熟考的思考)」を模倣したものであり、数学、論理パズル、プログラミングといった、従来の小型モデルが苦手としていたタスクにおいて劇的な性能向上をもたらした。

例えば、**Ministral 3 14B Reasoning**モデルは、難関数学ベンチマークであるAIME '25において**85%**の正答率を記録している¹³。これは、通常であれば数百億～数千億パラメータを持つモデルでなければ達成困難なスコアであり、14Bという軽量モデルが「長く考える」ことで物理的なパラメータの制約を超越できることを実証している¹³。この機能により、エッジデバイス上でも、通信を介さずに高度な数学的計算や論理的判断が可能となる。

4.3 エージェントティック・ワークフローとツール利用

Ministral 3は、ツール(Tools)やAPIを自律的に呼び出す能力(Function Calling)が強化されており、エージェントシステムの構成要素として最適化されている¹⁰。具体的には、ユーザーの曖昧な指示を解釈し、必要な外部ツール(検索エンジン、計算機、カレンダー操作など)を選択し、適切な引数で実行する能力を持つ。

また、視覚情報(画像)を入力として受け取り、その内容に基づいてアクションを起こすことも可能で

あるため、ロボットの目としての活用や、スクリーン上の情報を読み取って操作する自動化エージェントとしての応用が期待されている¹²。

5. ベンチマーク評価と競合比較分析

Mistral 3ファミリーは、ベンチマークにおいて極めて高い競争力を示している。ここでは、主要な指標に基づき、競合モデル(Llama 4、GPT-4o、Claude等)との比較分析を行う。

5.1 Mistral Large 3 vs. Llama 3.1 405B / GPT-4o

Mistral Large 3は、オープンソース(Open Weights)モデルの最高峰を目指しており、LMArenaリーダーボードではオープンソース非推論モデルカテゴリで2位(全体6位)にランクインしている¹³。

主要ベンチマーク比較(推定値および提供資料に基づく統合分析)

ベンチマークカテゴリ	指標	Mistral Large 3 (675B)	Llama 3.1 405B	GPT-4o	比較分析
総合知識	MMLU	85.5 ¹⁰	86.0 ¹⁸	88.7 ¹⁹	GPT-4oには及ばないものの、Llama 3.1 405Bとほぼ同等(誤差範囲内)の知識レベルを有している。
推論・専門知識	GPQA Diamond	43.9 ¹⁰	50.5-57.2 ²⁰	66.3 ²⁰	超難問タスクではGPT-4oやLlama 4 Maverickに差をつけられているが、実用域では高い性能。

コーディング	HumanEval / LiveCodeBench	34.4 (LCB) ¹⁰	43.4 (LCB) ²¹	43.2 (LCB) ²²	コーディング特化モデル (Codestral 等)には劣るが、汎用モデルとしては十分な水準。
数学	MATH	61.2% (vs) ²³	61.2%以上	76.6% (Est)	数学分野ではクローズドモデルや Llama 4 Maverickの方が優位な傾向にある。

10

分析とインサイト:

Mistral Large 3の強みは、**「多言語能力」と「推論効率」**のバランスにある。MMLUスコアが示す通り、一般的な知識量では世界トップクラスのモデルと遜色ない。一方で、GPQAやLiveCodeBenchなどの高難易度タスクでは、Llama 4 MaverickやGPT-4oといった最新世代モデル(2025年時点のSOTA)に対して若干の後れを取っているデータも見られる¹⁰。しかし、Mistral Large 3が41Bアクティブパラメータで動作することを考慮すれば、そのコストパフォーマンスは圧倒的である。Llama 3.1 405Bのようなデンスモデルを運用する場合と比較して、メモリ要件と電力消費を大幅に削減できるため、企業が自社環境でホスティングする際の現実的な最適解となり得る。

5.2 Ministral 3 vs. Llama 3.2 / Gemma 3

エッジ向けモデルの比較では、Ministral 3(特に14B)が独自の地位を築いている。

モデル	パラメータ	MMLU	MATH	特記事項
Ministral 3 14B	14B	82.0 ¹⁰	90.4 ¹⁰	Reasoningバリエーションの数学性能が突出している。
Mistral Small	24B	80.7 ²³	46.0 ²³	旧世代。

3.2				Ministral 14Bの方が軽量かつ高性能な場合がある。
Llama 3.2 3B	3B	63.4 ²⁴	48.0 ²⁴	Ministral 3 3Bと競合。軽量だが推論能力は限定的。

10

Ministral 3 14Bは、ベンチマーク上では旧世代のMistral Small 3.2(24B)を多くの指標で上回っており、パラメータ数が少ないにもかかわらず高性能化に成功している²³。特にReasoningモデルの数学性能(MATH 90.4%)は驚異的であり、エッジデバイスで動作するモデルとしては異次元のスコアである。

ただし、ユーザーレビューでは、Ministral 3が長文生成時に「ループ(繰り返し)」に陥りやすいという不安定さも報告されており、ファインチューニングやサンプリングパラメータの調整が必要な場合があることが示唆されている²⁵。

5.3 日本語および多言語性能

Mistral 3ファミリーは、英語中心のモデルとは一線を画す多言語対応能力を持つ。40以上の言語に対応し、事前学習段階から多言語データを含めているため、トークナイザーの効率も各言語に最適化されている⁵。

日本語性能に関しては、Mistral Small v3(関連モデル)の評価においてMMLU(知識)で約80%のスコアが推定されており、これは70Bクラスのモデルに匹敵する水準である²⁶。これは、日本のユーザーにとっても、翻訳レイヤーを挟まずにネイティブな日本語で高度な推論や要約を行えることを意味しており、特にニュアンスの理解が求められるビジネス文書や創作活動において有用性が高い。

6. オープンソース戦略とエコシステムへの影響

Mistral AIの戦略的価値は、技術的性能だけでなく、その提供形態にある。**Apache 2.0**ライセンスの採用は、クローズド化が進むAI業界において強力な差別化要因となっている⁵。

6.1 ライセンスの自由と主権AI(Sovereign AI)

Llama 3.x以降の「Llama Community License」は商用利用を許可しているものの、月間アクティブユーザー数が7億人を超える企業には別途ライセンス契約を求めるなどの制約が存在する。対して、Mistral 3のApache 2.0ライセンスは、オープンソース定義(OSI)に準拠した非常に寛容なライセンスであり、利用者の規模や用途(商用、研究、内部利用)を問わず自由に使用、修正、再配布が可能である¹⁴。

この特徴は、データ主権 (Sovereignty) を重視する企業や国家にとって極めて重要である。例えば、金融大手 **HSBC** や自動車大手 **Stellantis** は、Mistral のモデルを採用し、自社のセキュアなインフラストラクチャ内で運用している²⁸。外部の API にデータを送信することなく、自社のデータを学習させた独自の「プライベート LLM」を構築できることは、セキュリティ要件の厳しい産業において必須条件であり、Mistral 3 はこの需要に完全に応えるソリューションとなっている。

6.2 開発者エコシステムとの統合

Mistral 3 は、リリースと同時に広範なエコシステムに統合された。

- 推論エンジン: vLLM や TensorRT-LLM といった主要なオープンソース推論ライブラリが「Day 0 (初日)」からサポートを提供している¹³。これにより、エンジニアは複雑な統合作業なしに、pip install レベルの手軽さで最新モデルを試すことができる。
- クラウドプラットフォーム: AWS Bedrock, Azure Foundry, IBM WatsonX, Google Cloud など、主要なパブリッククラウドすべてで利用可能である⁸。これにより、企業は既存のクラウド契約の枠組みの中で Mistral 3 を利用できる。
- 量子化とデプロイ: Hugging Face 上では、FP8 や NVFP4 といった様々な量子化フォーマットのチェックポイントが公開されており、リソースに制約のある環境でも即座にデプロイが可能となっている⁶。

7. 経済性と運用コストの分析

「効率的な規模 (Efficient Scale)」を掲げる Mistral 3 は、経済性の面でも優位性を持つ。

7.1 トークンエコノミクスと推論コスト

Mistral Large 3 の API 価格は、入力 100 万トークンあたり 0.50 ドル、出力 100 万トークンあたり 1.50 ドルと設定されている (スニペット 10 に基づく)。これは、同等クラスの性能を持つ旧世代モデルや一部の競合モデルと比較して極めて競争力のある価格設定である¹⁰。

特に、RAG アプリケーションでは大量のコンテキストを入力する必要があるため、安価な入力トークンコストと 256K のウィンドウは、運用コスト (TCO) の大幅な削減に直結する。

7.2 エッジ運用のコストメリット

Mistral 3 を活用することで、クラウドへの API リクエスト自体を削減できる。簡単なタスクや個人情報の処理をローカルの Mistral 3B/8B で処理し、高度な推論が必要な場合のみクラウドの Mistral Large 3 にエスカレーションする「ハイブリッド AI アーキテクチャ」を構築することで、通信コストと API コストを最小化しつつ、全体のシステム性能を維持することが可能となる¹⁷。

8. 結論: AI 民主化の新たな標準

2025 年 12 月 2 日に発表された Mistral 3 は、単なる高性能モデルのリリースにとどまらず、AI 技術の

「民主化」と「実用化」を加速させる触媒としての役割を果たしている。

1. **MoEの勝利**: 675Bパラメータの知識を41Bのコストで運用可能にしたMistral Large 3は、MoE技術が実用段階の頂点に達したことを証明した。
2. **ハードウェアとの融合**: NVIDIA GB200/Blackwellとの共進化により、ソフトウェアとハードウェアの境界線が溶解し、かつてない推論効率が実現された。
3. **エッジへの知能の拡散**: Mistral 3とReasoning機能により、スーパーコンピュータ並みの推論能力が、手のひらの上のデバイスで利用可能になりつつある。
4. **真のオープン**: Apache 2.0ライセンスの堅持は、特定巨大企業によるAI技術の独占を防ぎ、世界中の開発者が自らの手でAI未来 (AI Future) を所有することを可能にした¹¹。

Mistral 3は、AI開発の主戦場が「パラメータ数の競争」から「効率、適用性、そして開放性」へと移行したことを象徴している。企業、研究者、そして個人の開発者は、この強力なツールセットを手にすることで、AIアプリケーションの可能性を無限に広げることができるだろう。Mistral AIが提示した「分散型インテリジェンス」のビジョンは、2026年以降のAI社会の基盤となる標準アーキテクチャを示唆していると言える。

免責事項: 本報告書は、2025年12月時点での公開情報および提供された資料に基づき作成されています。ベンチマークスコアや仕様は、ハードウェア構成や測定条件により変動する可能性があります。

引用文献

1. NVIDIA & Mistral AI Partner Up, Collab Will Accelerate New Family of Open Models, 12月 6, 2025にアクセス、
<https://www.techpowerup.com/343621/nvidia-mistral-ai-partner-up-collab-will-accelerate-new-family-of-open-models>
2. NVIDIA Partners With Mistral AI on Open Models, 12月 6, 2025にアクセス、
<https://insidehpc.com/2025/12/nvidia-partners-with-mistral-ai-on-new-open-models/>
3. NVIDIA Partners with Mistral AI to Accelerate New Family of Open Models, 12月 6, 2025にアクセス、
<https://www.hpcwire.com/aiwire/2025/12/03/nvidia-partners-with-mistral-ai-to-accelerate-new-family-of-open-models/>
4. Mistral's latest open-source release bets on smaller models over large ones - here's why, 12月 6, 2025にアクセス、
<https://www.zdnet.com/article/mistrals-latest-open-source-release-says-smaller-models-beat-large-ones-heres-why/>
5. Mistral 3 : Best Open-sourced model is here !! | by Mehul Gupta | Data Science in Your Pocket - Medium, 12月 6, 2025にアクセス、
<https://medium.com/data-science-in-your-pocket/mistral-3-best-open-sourced-model-is-here-3b93a6b2b2e8>
6. mistralai/Mistral-Large-3-675B-Instruct-2512-NVFP4 - Hugging Face, 12月 6, 2025にアクセス、

- <https://huggingface.co/mistralai/Mistral-Large-3-675B-Instruct-2512-NVFP4>
7. NVIDIA-Accelerated Mistral 3 Open Models Deliver Efficiency, Accuracy at Any Scale, 12月 6, 2025にアクセス、
<https://developer.nvidia.com/blog/nvidia-accelerated-mistral-3-open-models-deliver-efficiency-accuracy-at-any-scale/>
 8. Mistral Large 3 now available on IBM watsonx, 12月 6, 2025にアクセス、
<https://www.ibm.com/new/announcements/mistral-large-3-now-available-on-ibm-watsonx>
 9. mistralai/Mistral-Large-3-675B-Instruct-2512 - Hugging Face, 12月 6, 2025にアクセス、
<https://huggingface.co/mistralai/Mistral-Large-3-675B-Instruct-2512>
 10. Mistral 3 Review: 5 Insane Benchmarks & Local Setup Secrets, 12月 6, 2025にアクセス、
<https://binaryverseai.com/mistral-3-review-benchmarks-api-pricing-install/>
 11. French AI shop Mistral rolls out full suite of Apache-licensed models, 12月 6, 2025にアクセス、
https://www.theregister.com/2025/12/02/mistral_3/
 12. Mistral Large 3 and Ministral 3 family now available first on Amazon Bedrock - AWS, 12月 6, 2025にアクセス、
<https://aws.amazon.com/about-aws/whats-new/2025/12/mistral-large-3-ministral-3-family-available-amazon-bedrock/>
 13. Introducing Mistral 3, 12月 6, 2025にアクセス、
<https://mistral.ai/news/mistral-3>
 14. Models - from cloud to edge - Mistral AI, 12月 6, 2025にアクセス、
<https://mistral.ai/models>
 15. Reasoning - Mistral AI Docs, 12月 6, 2025にアクセス、
<https://docs.mistral.ai/capabilities/reasoning>
 16. When you prompt a non-thinking model to think, does it actually improve output? - Reddit, 12月 6, 2025にアクセス、
https://www.reddit.com/r/LocalLLaMA/comments/1jnehr4/when_you_prompt_a_nonthinking_model_to_think_does/
 17. Un Ministral, des Ministraux - Mistral AI, 12月 6, 2025にアクセス、
<https://mistral.ai/news/ministraux>
 18. LLM overkill is real: I analyzed 12 benchmarks to find the right-sized model for each use case 🤖 : r/LocalLLaMA - Reddit, 12月 6, 2025にアクセス、
https://www.reddit.com/r/LocalLLaMA/comments/1qlscfk/llm_overkill_is_real_i_analyzed_12_benchmarks_to/
 19. The AI Heavyweight Bout: GPT-4o vs Claude 3.5 vs Mistral Large 2 vs Llama 3.1 - Medium, 12月 6, 2025にアクセス、
<https://medium.com/@cognidownunder/the-ai-heavyweight-bout-gpt-4o-vs-claude-3-5-vs-mistral-large-2-vs-llama-3-1-6be5df8ecf93>
 20. LLM Leaderboard | Compare Top AI Models for 2024 - YourGPT, 12月 6, 2025にアクセス、
<https://yourgpt.ai/tools/llm-comparison-and-leaderboard>
 21. o1 vs Llama 4 Maverick - LLM Stats, 12月 6, 2025にアクセス、
<https://llm-stats.com/models/compare/o1-2024-12-17-vs-llama-4-maverick>
 22. Model Statistics - CodingFleet, 12月 6, 2025にアクセス、
https://codingfleet.com/models?sort=livebench_coding&period=all
 23. Llama 4 Maverick vs Mistral Small 3 24B Base, 12月 6, 2025にアクセス、
<https://llm-stats.com/models/compare/llama-4-maverick-vs-mistral-small-24b-b>

[ase-2501](#)

24. Llama 3.2 3B Instruct vs Ministral 8B Instruct - LLM Stats, 12月 6, 2025にアクセス、
<https://llm-stats.com/models/compare/llama-3.2-3b-instruct-vs-ministral-8b-instruct-2410>
25. Mistral 3 Blog post : r/LocalLLaMA - Reddit, 12月 6, 2025にアクセス、
https://www.reddit.com/r/LocalLLaMA/comments/1pcayfs/mistral_3_blog_post/
26. 日本語対応 ! Mistral Small v3 解説 - (株)Qualiteg, 12月 6, 2025にアクセス、
https://blog.qualiteg.com/mistral_small_v3_introduction/
27. 🐱 Everything Amazon launched at re:Invent 2025, 12月 6, 2025にアクセス、
<https://www.theneurondaily.com/p/everything-amazon-launched-at-re-invent-2025>
28. Mistral Launches 3 Models that Land HSBC, Nvidia Backing, 12月 6, 2025にアクセス、
<https://www.pymnts.com/artificial-intelligence-2/2025/mistral-launches-3-models-that-land-hsbc-nvidia-backing>
29. Run Mistral Large 3 & Ministral 3 on vLLM with Red Hat AI on Day 0: A step-by-step guide, 12月 6, 2025にアクセス、
<https://developers.redhat.com/articles/2025/12/02/run-mistral-large-3-ministral-3-vllm-red-hat-ai>
30. Mistral Large 3, 12月 6, 2025にアクセス、
<https://docs.mistral.ai/models/mistral-large-3-25-12>