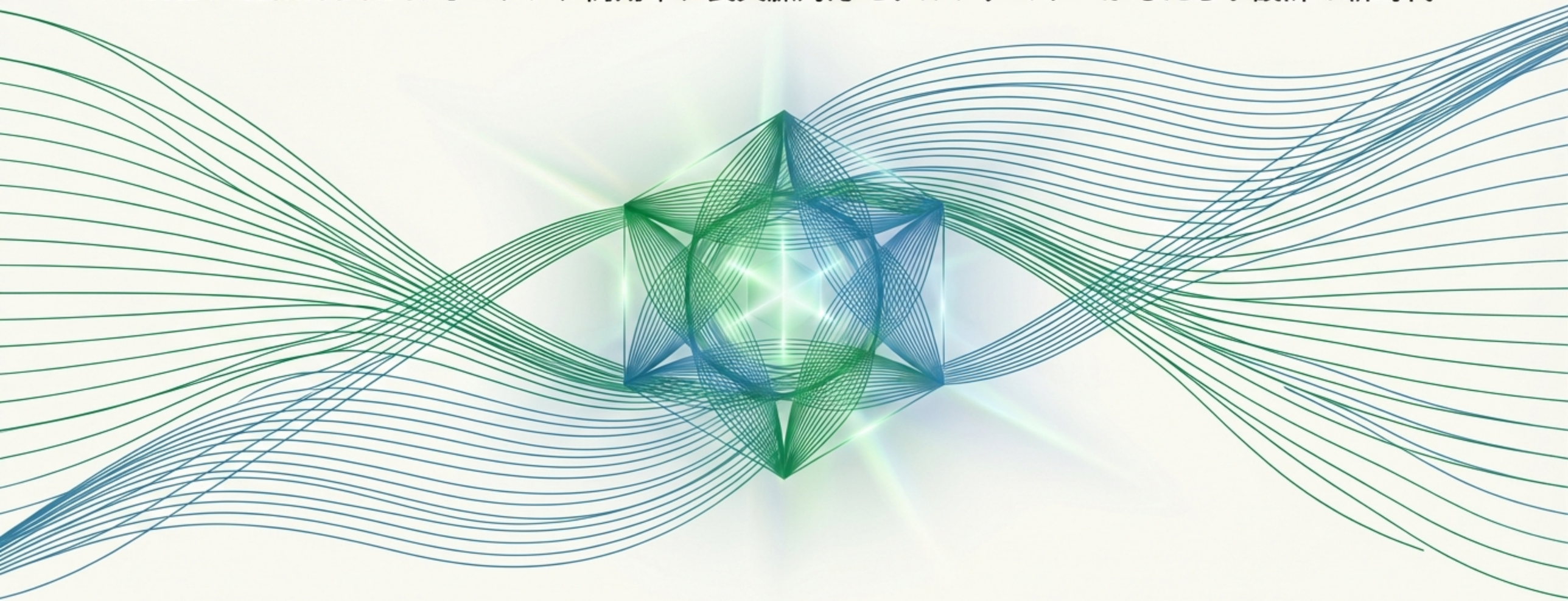
AIエージェント開発の、技術的なターニングポイント

NVIDIA Nemotron 3: オープン、高効率、長文脈対応モデルファミリーがもたらす設計の新時代



性能・効率・開放性を同時に実現し、次世代AIエージェントの構築を加速する。

現代のAIエージェントシステムが直面する設計上の課題

現代のエージェントシステムは、複数の専門エージェント（検索、計画、ツール実行、検証）が連携して動作します。この実現には、従来のモデルでは両立が困難だった複数の要件が求められます。

高速スループット (Fast Throughput)

多数のエージェントを並列で低遅延に実行する必要性。

大規模コンテキストでの一貫性 (Coherence over Large Contexts)

長大な文書や会話履歴全体を矛盾なく処理する能力。

高度な推論精度 (Strong Reasoning Accuracy)

複数ステップにわたる複雑な計画やツール呼び出しを正確に実行する。

カスタマイズと拡張性 (Customization and Extensibility)

特定ドメインへの特化や、自由なデプロイを可能にするオープン性。

その答えが Nemotron 3: エージェントAI時代のためのオープンモデルファミリー

NVIDIA Nemotron 3は、Nano、Super、Ultraの3モデルで構成されるモデル ファミリーです。エージェントAIの特有の要求に応えるため、性能、効率、そして開発者のためのオープン性を徹底的に追求して設計されました。

Nemotron 3 Nano

Purpose:
高スループット、高効率な推論

Status:
本日より利用可能

Nemotron 3 Super

Purpose:
協調型エージェント、大規模ワークロード

Status:
2026年前半に登場予定

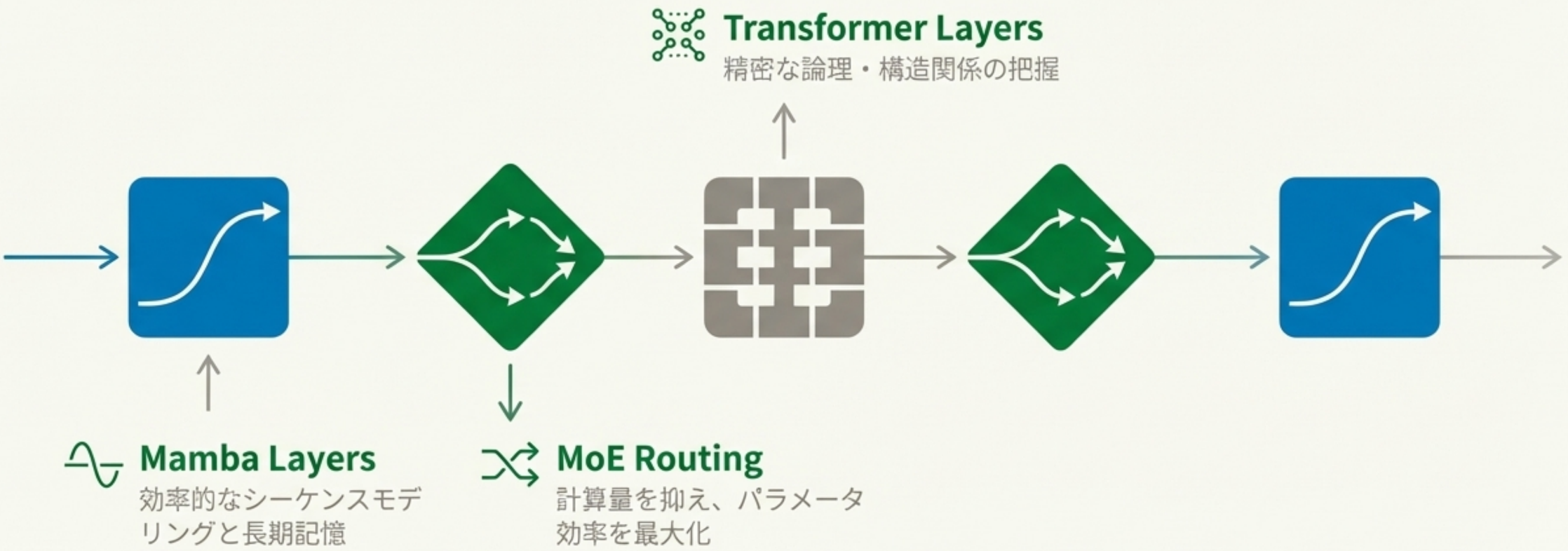
Nemotron 3 Ultra

Purpose:
最先端の精度と複雑な推論タスク

Status:
2026年前半に登場予定

Nemotron 3の心臓部: Hybrid Mamba-Transformer MoE アーキテクチャ

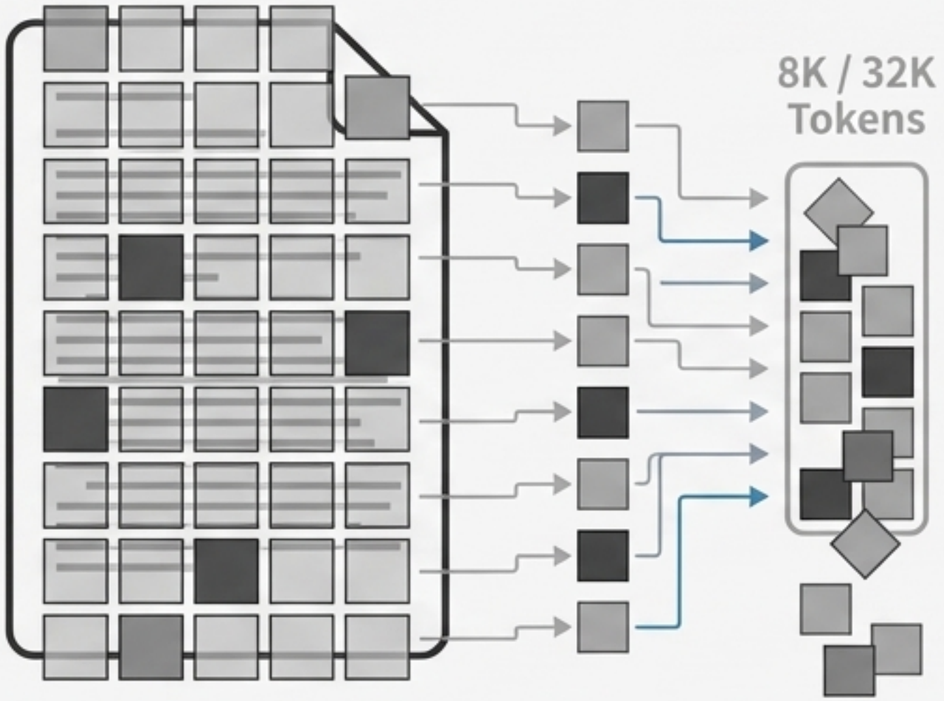
3つのアーキテクチャを単一のバックボーンに統合。それぞれの長所を組み合わせることで、推論スループ-プットを最大化しつつ、最先端の精度を維持します。



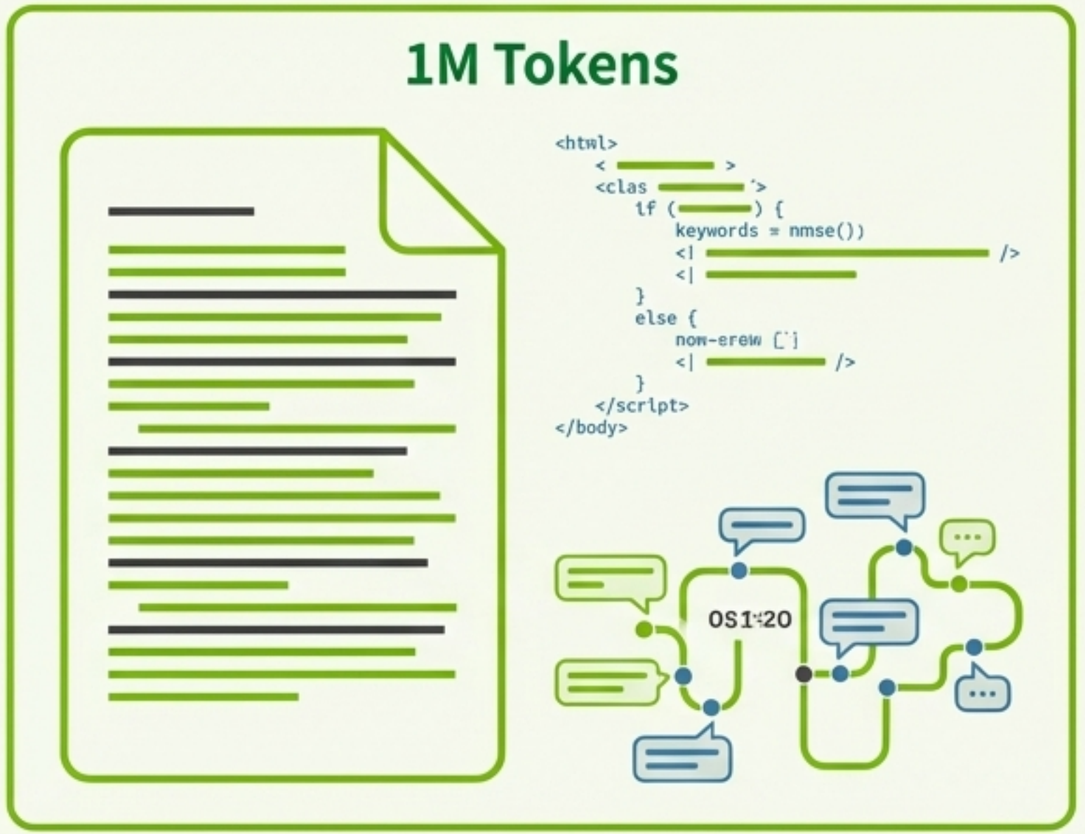
1Mトークンコンテキスト: ワークフローの設計が単純化される

巨大なコードベース、複数の長文ドキュメント、長時間の会話履歴を、断片化することなく単一のコンテキストウィンドウで処理。RAGなどのパイプラインで、複雑なチャンキング戦略が不要になります。

従来の手法

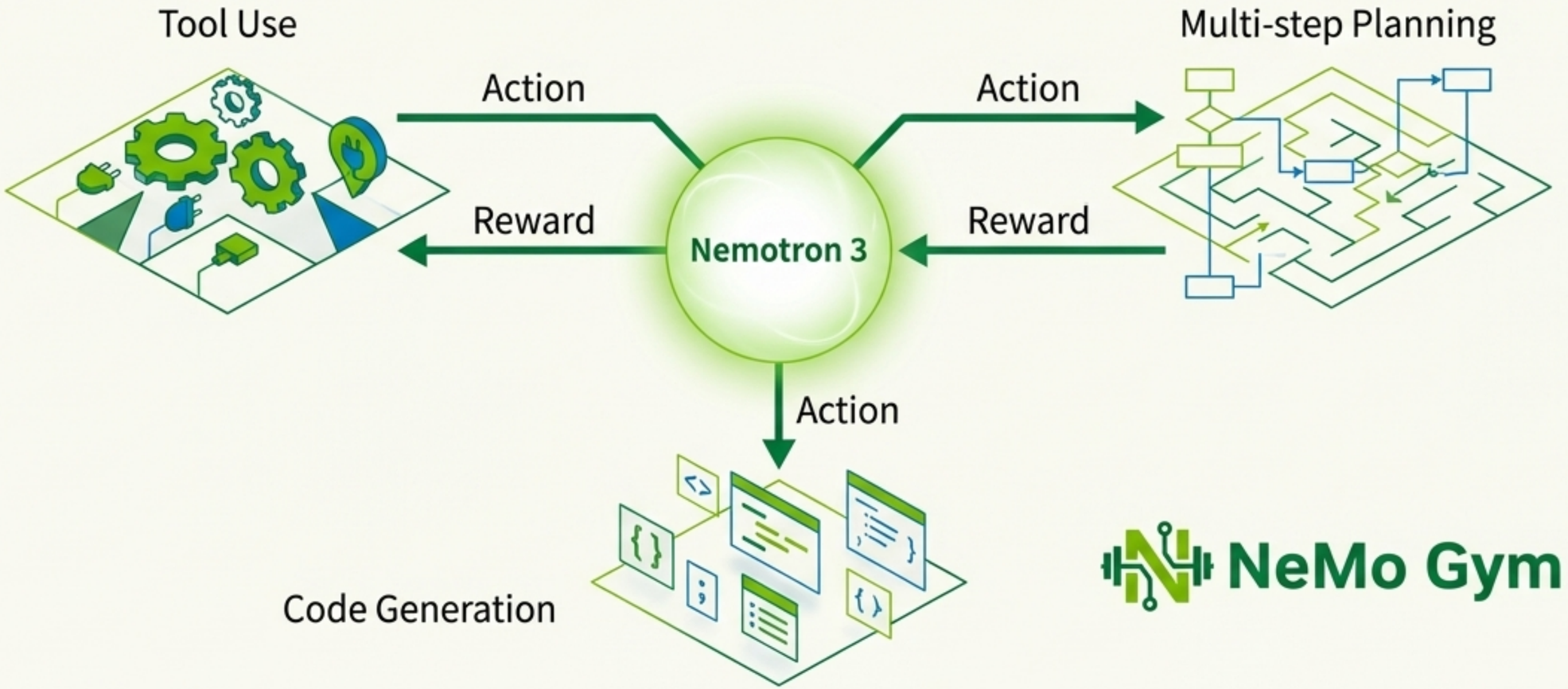


Nemotron 3



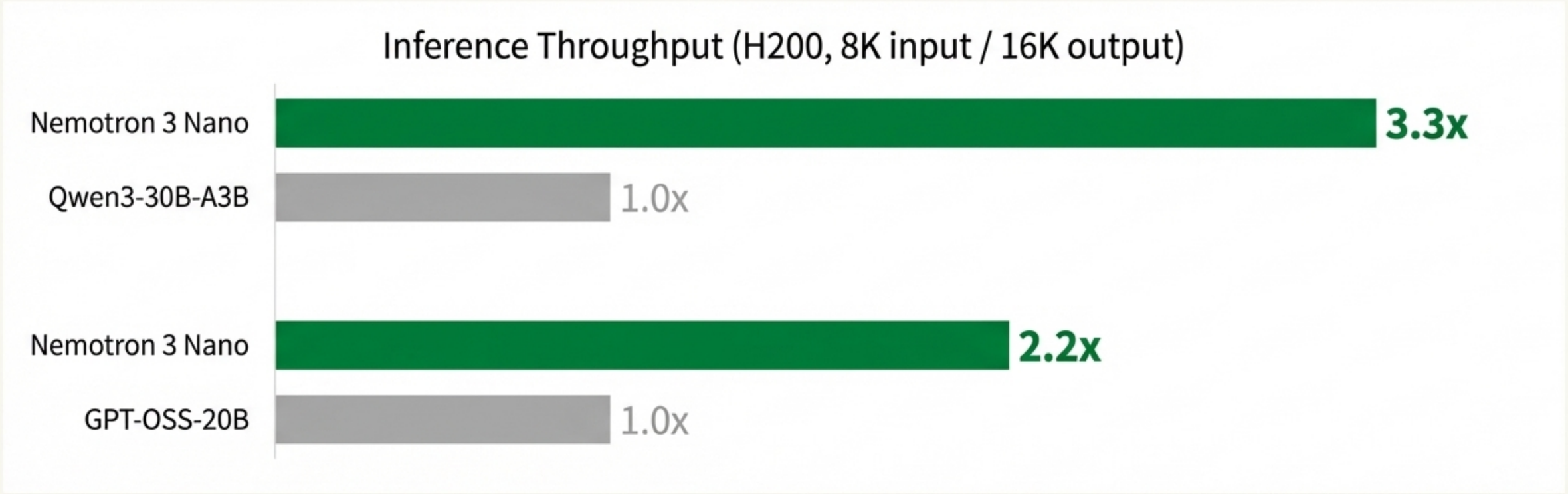
実世界のタスクへのアライメント: NeMo Gymによるマルチ環境強化学習

単一の応答だけでなく、複数ステップにわたるタスク（ツール呼び出し、コード生成、計画立案）の成功を評価。NeMo Gymはオープンソースであり、開発者は独自の環境を構築してモデルを特定ドメインにさらに特化させることが可能です。



圧倒的なスループット効率を実現

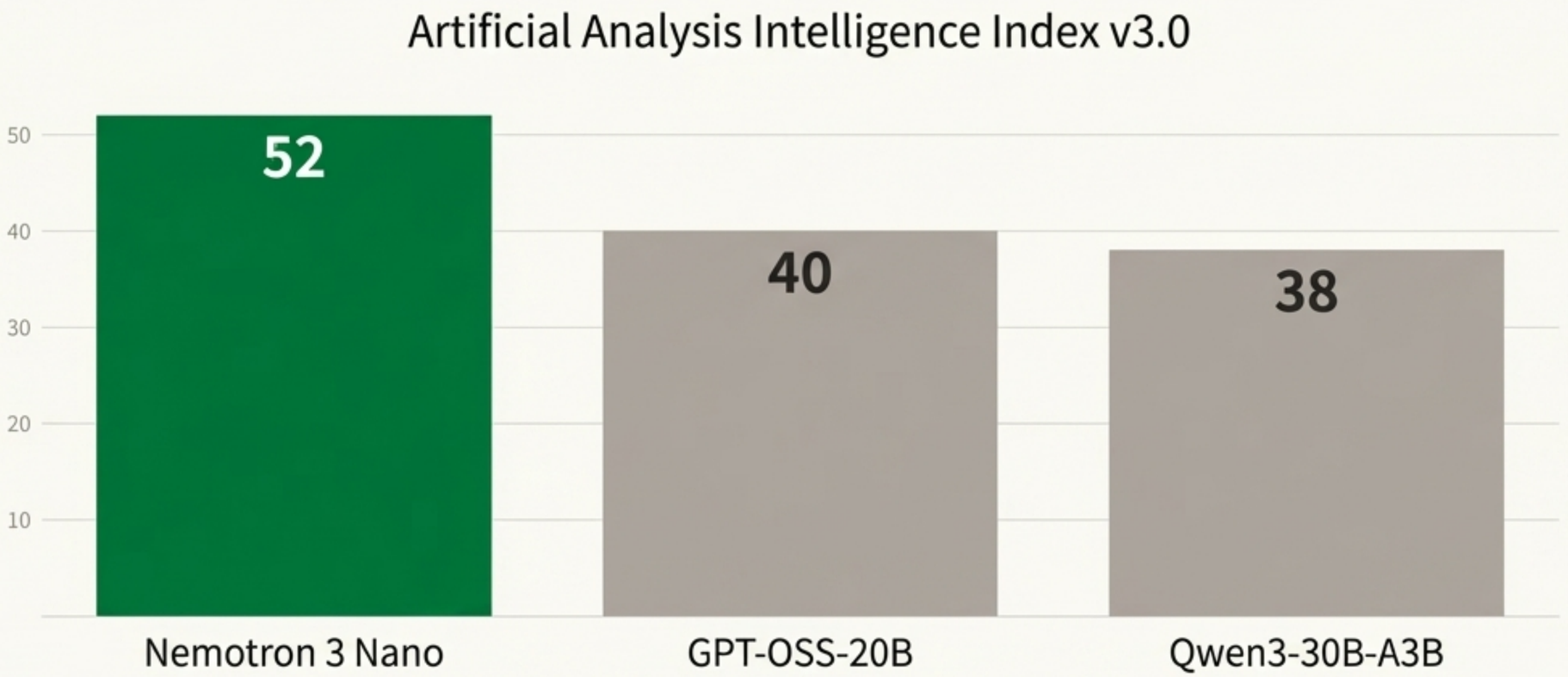
Nemotron 3 Nanoは、同規模のモデルと比較して推論スループットで他を凌駕します。



Key Insight: Hybrid MoEアーキテクチャがもたらす、コスト効率の劇的な向上。

クラス最高水準の精度と推論能力

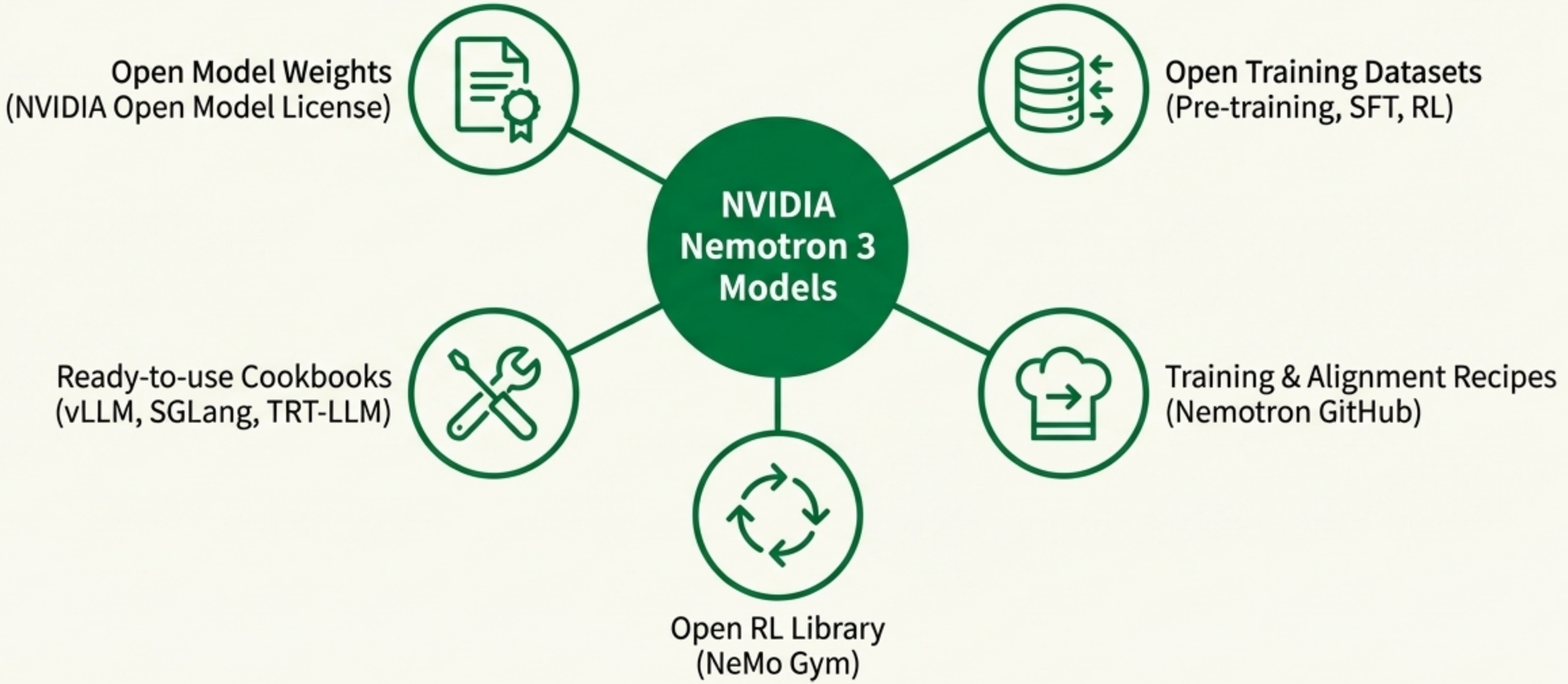
Nemotron 3 Nanoは、一般的なベンチマークにおいて主要な同規模モデルを上回るスコアを達成。



Key Insight: 高度な強化学習とデータセットにより、効率と精度を両立。

モデルだけではない: 完全なオープン開発エコシステム

Nemotron 3は、モデル、データ、ツール、レシピを統合したプラットフォームです。これにより、開発者はモデルの動作を完全に理解し、再現し、そして自由にカスタマイズできます。



前例のない透明性: オープンな学習データセット

高性能で信頼性の高いモデルがどのように構築されるか。その全貌を理解するために、モデル開発の各段階で使用されたデータセットを公開します。

Pre-training Data

Nemotron-pretraining: 3兆トークン。コード、数学、推論に豊富なカバレッジ。

Post-training Data

Nemotron-post-training 3.0: 1300万サンプル。SFTとRLのためのコーパス。

Reinforcement Learning Data

Nemotron-RL datasets: ツール使用、計画、マルチステップ推論のための厳選されたデータセットと環境。

Safety Data

Nemotron agentic safety dataset: 約1万1000件のエージェントワークフロートレース。安全性とセキュリティリスクの評価・軽減用。

開発者へのインパクト: 設計自由度が跳ね上がる

Nemotron 3の技術は、単なる性能向上以上の価値を提供します。これまで複雑だったエージェントワークフローの設計を根本から変え、新しい可能性を拓きます。

1M Context & Hybrid MoE

巨大コンテキスト対応により
RAG／長文情報統合ワークフローの設計が単純化

High Throughput & MoE

高スループットでの
並列エージェント運用が低コスト化

Open Weights, Recipes, Data

完全オープンなスタックにより
カスタムモデル開発が容易化

今すぐ始める: NVIDIAクックブックと主要ツールで迅速に導入

すぐに使えるクックブックと、広く使われているツールへの対応により、数分でNemotron 3 Nanoを起動できます。

NVIDIA Cookbooks

vLLM

高スループットな
連続バッチ処理

SGLang

マルチエージェントの
ツール呼び出しに最適化

TensorRT-LLM

低遅延な本番環境向け

Broad Community Support

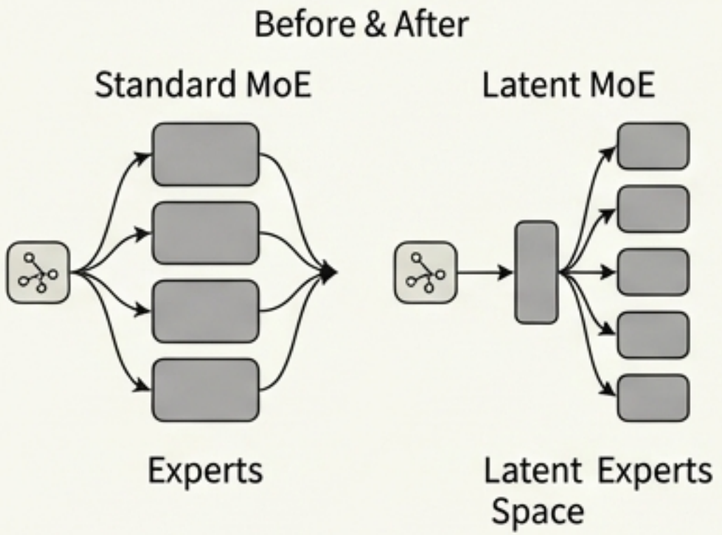
- Llama.cpp
- LM Studio
- Unsloth

GeForce RTXデスクトップからDGX Sparkまで、あらゆるNVIDIA GPUで利用可能。

Nemotron 3 Super & Ultraが拓く未来: さらなる精度と効率へ

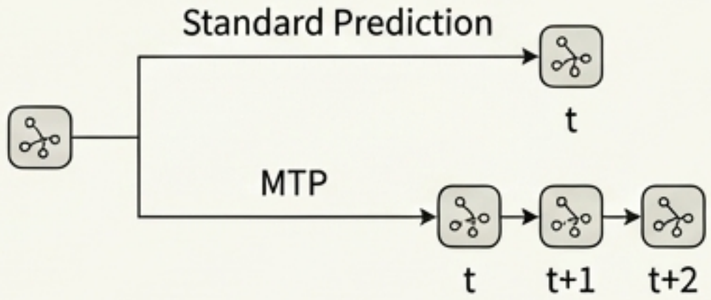
2026年前半に登場するSuperとUltraは、推論の深化と効率化をさらに
推し進めるための先進技術を搭載します。

Latent MoE



同じ推論コストで4倍の専門家（エキスパート）を活用し、専門性を向上。

Multi-Token Prediction (MTP)



複数の未来トークンを一度に予測し、スループットを大幅に向上。

NVFP4 Training

NVIDIAの4ビット浮動小数点形式により、学習と推論でクラス最高のコスト精度を達成。

オープンなモデルへの継続的なコミットメント

Nemotron 3は、透明性と開発者のエンパワーメントに対するNVIDIAの姿勢を明確に示すものです。モデルの実行、デプロイ、構築方法の調査、そして独自モデルの学習まで、すべてをNVIDIAのオープンなリソースで完結できます。

モデルウェイトを公開

学習レシピを公開

データセットを公開

あなたも、次世代の推論モデル構築へ

オープンな研究を加速することは、Nemotronチームの最優先事項です。Nemotronのオープンモデルとデータセットを活用し、推論性能の向上を目指す新しいコミュニティコンペティションに、ぜひご参加ください。

Call to Action

- Nemotron Model Reasoning Challengeに登録
- Hugging Faceでモデルとデータセットを探索
- Nemotronデベロッパーページで詳細を確認
- DiscordのNemotronチャンネルに参加