エージェント型 AI のための、
オープンで効率的なマルチモーダル モデル。

NVIDIA Nemotron

NVIDIA Nemotron とは?

NVIDIA Nemotron™ は、長期稼働する専門的なエージェント型 AI システム向けに構築された、高効率かつオープンなマルチモーダル モデル、データセット、テクノロジーからなるファミリーです。 高度な推論、コーディング、視覚理解、安全性、音声、情報検索向けに設計された Nemotron モデルは、エージェントが複雑なタスクを優れた精度で、低コストで迅速に完了できるように支援します。

他のオープンモデル（Llama等）との決定的違い

合成データ生成（蒸留）の公式許可

Llama 3等多くのモデルは、出力結果を用いて他モデルを学習させることを規約で制限しています。一方、Nemotronはそれを許容・推奨しており、自社の専門用語を反映した**高品質な学習データを合法的に大量生成**し、ローカルモデルの構築に利用できます。

エンタープライズ特化のアライメント

「汎用AI」ではなく「商用AI」として設計。出力品質を評価する強力なRewardモデル（Nemotron-4 340B等）が提供されており、ハルシネーション（もっともらしい嘘）を抑え、企業のコンプライアンスに準拠したAIを構築可能です。

ターゲット業界と活用シナリオ

医療・創薬

患者の電子カルテや新薬データなど、外部クラウドに出せない機密情報を扱う用途。閉域網内でのローカルエージェントとして稼働。

金融・法務

膨大な社外秘の契約書や顧客データを読み込ませたRAG（検索拡張生成）基盤。セキュリティを担保しながらコンプライアンス業務を自動化。

製造・EDA

製造現場のエッジ環境における自律AIや、電子設計自動化（EDA）ツールへの組み込み。低遅延・高セキュリティ環境で真価を発揮。

Nemotronの3つの特徴

高いカスタマイズ性

NVIDIA NeMoフレームワークとシームレスに連携。企業の機密データを用いたファインチューニングやRAG（検索拡張生成）の構築が容易です。

圧倒的な効率と速度

一部モデルではMoE（Mixture of Experts）アーキテクチャを採用。巨大な知識を持ちながら、推論時は計算リソースを最適化し高速処理を実現します。

エンタープライズ品質

RLHFなどの技術により、安全性とアライメントを徹底。商用利用を前提としたコンプライアンス基準を満たす、信頼性の高い出力を行います。

エージェント型 AI のための、
オープンで効率的なマルチモーダル モデル。

NVIDIA Nemotron

各モデルの推奨VRAM・ハードウェア要件

VRAM容量はコンテキスト長により変動します。

モデル名 / 規模	VRAM (16-bit)	VRAM (4-bit時)	推奨GPU構成例 (4-bit量子化時)
Nemotron-3 Nano 316億パラメーター	約 73 GB	約 18 GB	・RTX 5090 (32GB) × 1基 ・RTX PRO4500 Blackwell (32GB) × 1基
Nemotron-3 Super 1,200億パラメーター	約 276 GB	約 69 GB	・RTX PRO6000 Blackwell (96GB) × 1基 ・DGX Spark GB10 (128GB)
Nemotron-3 Ultra 5,500億パラメーター	約 1.27 TB	約 317 GB	・H200 NVL (141GB) × 4基 ・XpertStation WS300 (748GB)

ubuntu

NVIDIA GEFORCE RTX

AMD THREADRIPPER PRO

RTX PRO6000 96GB搭載モデル
メモリ/ストレージのカスタマイズも可能

■CPU： Ryzen Threadripper PRO 9965WX 4.25GHz-5.4GHz/24C/48T
■チップセット：AMD WRX90
■CPUクーラー：360mm水冷
■メモリ：128GB（16GB x8）DDR5-5600
■SSD ： 1TB M.2 NVMe-SSD
■OS： Ubuntu 24.04 LTS インストール代行
■GPU：NVIDIA RTX PRO 6000 Blackwell Max-Q 96GB-GDDR7
■電源：1,200W/100V 80 Plus Platinum 認証
■LAN：10GbE×2
■3年間センドバック保証

APPLEID
Nemotron推奨モデル

価格はお問い合わせ下さい

アプライド Computer&Internet Audio&Visual
アプライド株式会社
広域システム営業部