開発を加速し、エンタープライズ AI を解き放つ。SUPERMICRO と NVIDIA の RAG 対応インフラストラクチャ

概要

人工知能（AI）の急速な進化により、企業は大規模言語モデル（LLM）をより安全かつ効果的に活用し、ビジネスにインパクトを与える方法を模索しています。しかしながら、組織は既存と新規インフラの複雑さ、人員不足、正確で文脈に即した安全な AI 出力の提供といった、根深い課題に直面しています。

このソリューション概要では、エンタープライズ対応 AI 実現の鍵となる検索拡張生成（RAG: Retrieval-Augmented Generation）に焦点を当てます。RAG は、LLM を組織のデータに直接接続することで、厳格なデータガバナンスとプライバシーを維持しながら、正確でビジネスに特化した結果をもたらします。Supermicro の柔軟なシステムと NVIDIA の業界をリードする GPU を活用した AI ファクトリー・ソリューションは、高度な RAG パイプラインを導入するための、即時利用可能で拡張性のある基盤を提供します。事前トレーニング済みモデルと、エンタープライズナレッジ検索、堅牢なセキュリティ制御を組み合わせることで、企業は AI の導入を加速し、信頼性の高い成果を確保し、機密情報を確実に保護しながら、AI の可能性を最大限に引き出すことができます。

エンタープライズ AI 導入の障壁

企業は AI 導入において、数多くの実用的かつ技術的な課題に直面しています。日常的な IT および AI タスクの人員不足に加え、電力、冷却、設置スペースの制約により、今日の要求の厳しいワークロードに必要なインフラストラクチャの構築と維持が困難になっています。また、多くの組織は AI ハードウェアの導入経験が浅く、複雑なシステムを設計、導入、管理するための社内専門知識が不足しているため、ターンキー型の検証済みソリューションの必要性が高まっています。

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

インフラストラクチャ以外にも、事前学習済みの AI モデルは新たな障害をもたらします。これらのモデルは導入を迅速に進める手段となる一方で、不正確さや、「幻覚的な」応答を生成する可能性があり、ビジネスに関連する正確な回答を提供できないことも少なくありません。特にオープンソースモデルやクラウドソリューションを使用する場合、企業は機密情報を保護し、ガバナンス要件を満たす必要があるため、データのセキュリティと安全性は大きな懸念事項です。企業固有のデータを用いて基盤モデルを一から学習させるには、多くの場合、膨大な費用がかかるため、多くの組織はファインチューニングに頼らざるを得ません。しかし、それだけでは、完全に正確でコンテキストに基づいた安全な結果が得られない可能性があります。

これらの技術的およびセキュリティ上の課題は、AI 投資と導入成功のギャップに直接的な影響を与えています。マッキンゼーのレポートによると、企業の 92% が今後 3 年間で AI 投資を増やす計画であるにもかかわらず、自社の AI 導入が成熟していると考えるリーダーはわずか 1% にとどまっています。また、多くの企業はデジタル予算のごく一部を AI に割り当てており、58% の企業は AI への投資額が全体の 10% 未満にとどまっていると回答しています。これらの数字は、企業が AI 投資を成功に導き、大きな効果をもたらす導入を実現するためには、安全で拡張性に優れ、検証済みの AI ソリューションが喫緊に必要であることを浮き彫りにしています。

1 AI in the workplace: A report for 2025 | McKinsey

専用インフラストラクチャによるエンタープライズ AI の活用

人工知能（AI）の急速な進化は、エンタープライズ IT に根本的な変化をもたらし、あらゆる業界の組織が新たな価値を引き出し、業務を効率化し、競争優位性を維持できるようにします。AI の導入が加速するにつれ、企業は高度な AI 機能をイノベーションだけでなく、ビジネス変革の中核的な推進力として活用する方法を模索しています。

AI ファクトリー（専用設計で拡張性に優れたモジュール型インフラストラクチャ）は、この変革の基盤として台頭しています。最新の NVIDIA GPU を搭載し、Supermicro の柔軟でエンタープライズ対応のシステムを通じて提供される AI ファクトリーは、企業全体にわたる AI ワークロードの導入、管理、拡張のためのターンキーアプローチを提供します。

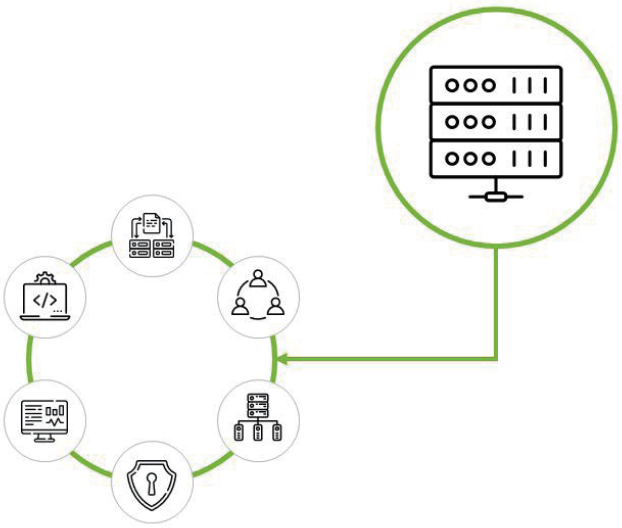
AI ファクトリーは魅力的な未来への道筋を示していますが、その可能性を最大限に実現するには、多くの企業が現在も直面している一連の実質的および技術的な課題を克服する必要があります。

事前学習済みモデルから、最適な情報提供や動作変更を自動的に行うコンテキストアウェアなインテリジェンスへ

事前学習済み AI モデルの限界を克服するために、企業はオープンソース AI と自社データを用いてモデルを事後学習させることができます。検索拡張生成（RAG）は、外部の知識と AI 推論を組み合わせることで、自社のビジネスコンテキストに合わせた、関連性が高く、正確で、安全な結果を提供し実現する、実用的な方法を提供します。

AI ファクトリーが企業にとって重要な理由

- AI 導入を容易にし、迅速な活用を実現
- 短期間でのシステム立ち上げ・運用開始
- エンタープライズ用途に求められる高い信頼性とセキュリティ
- インテリジェンスを大規模に生成・活用可能
- 生成 AI に必要な高度な計算リソースへの対応
- GPU 高密度かつ大容量データを扱う AI パイプラインに対応した IT データセンターへの進化



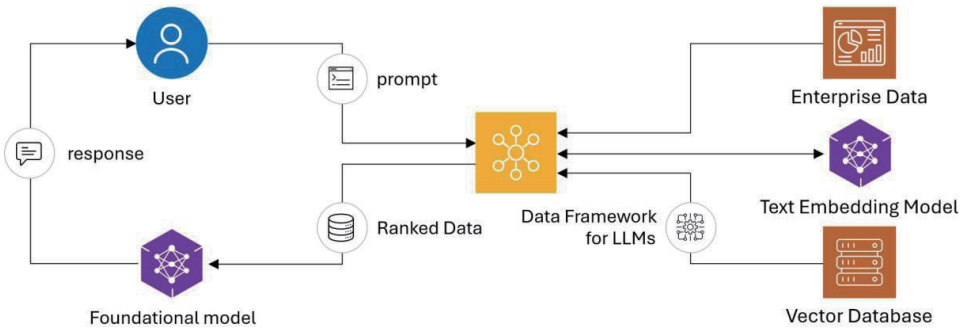
© 2026 Copyright Super Micro Computer, Inc. All rights reserved

RAG とは？

検索拡張生成（RAG）は、大規模言語モデル（LLM）を組織独自のデータに接続することで、AI が正確で、最新のビジネスコンテキストに適合した応答を生成できるようにします。RAG は、ナレッジベースやドキュメントリポジトリなどの社内ソースから関連情報を取得することで、モデル出力を実際の企業コンテンツに基づかせ、錯覚を軽減し、機密データを環境内で安全に維持します。このアプローチにより、企業は大規模モデルをゼロからトレーニングするコストや複雑さを伴わずに、より正確で信頼性の高い結果を得ることができます。

RAG の一般的な要件は次のとおりです。

- GPU：モデル推論および検索タスクの高速化。
- ストレージ：インデックス化された企業ドキュメントと埋め込みのホスティング。
- ソフトウェアスタック：事前トレーニング済みの LLM、検索システム（ベクターデータベースなど）、そしてそれらを接続するためのオーケストレーションツール。



戦略からソリューションへ

企業は、汎用的な AI を導入する際の制約を認識し、オープンソースモデルと自社データを組み合わせたトレーニングを行う戦略を採用するケースが増えています。このアプローチにより、企業はデータのプライバシーと関連性を維持しながら、AI の出力を自社のビジネスコンテキストに合わせて調整することが可能になります。

この戦略において中心的な役割を果たすのが、検索拡張生成（RAG）です。RAG は、事前トレーニング済みのモデルが実行時に企業の知識に動的にアクセスし、取り込むことを可能にします。しかし、RAG の潜在能力を最大限に引き出すには、強力で拡張性があり、複雑さや遅延を増大させることなく、すぐに導入できるインフラストラクチャが必要になります。

NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU を搭載した Supermicro RAG ソリューション

ここで Supermicro と NVIDIA が活躍します。両社は協力して、RAG ワークロードを導入直後から運用できるように設計した、完全な検証済みソリューションを提供します。NVIDIA の RAG ブループリントは、RAG パイプラインの構築と拡張のための実証済みのフレームワークを提供し、組織が自信を持って導入できるよう支援します。Supermicro は、この NVIDIA Blueprint を補完するために、RAG ワークロード向けに設計した NVIDIA-Certified Systems™（NVIDIA 認定システム）のポートフォリオを提供します。NVIDIA RTX PRO 6000 Blackwell Server Edition を含む NVIDIA の最新 GPU 向けに最適化された Supermicro のシステムは、リアルタイムの推論、検索、マルチモーダル AI アプリケーションに必要なパフォーマンス、拡張性、効率性を提供します。

次のセクションでは、エンタープライズ RAG 導入向けに最適化された Supermicro システムと NVIDIA GPU による、NVIDIA AI Enterprise ソフトウェアを活用した AI 開発と拡張を、加速、簡素化する方法について説明し、RAG 向け NVIDIA AI Blueprint にアクセスして導入を始める方法を説明します。

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

ステップ 1：Supermicro と NVIDIA の最適なインフラストラクチャを選択

企業全体で RAG を実現するための最初のステップは、それをサポートする適切なインフラストラクチャを選択することです。導入を成功させるには、性能、拡張性、ワークロード要件に適合したシステムを選択できることが重要です。Supermicro は、GPU 数、ネットワーク帯域幅、フォームファクターなど、RAG の要件を完全に満たすように設計・検証された、様々な種類の NVIDIA 認定システムを提供しています。それでは、RAG 導入戦略に最適化した Supermicro システムと NVIDIA GPU について見ていきましょう。

NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU に最適な Supermicro システムの構成

以下の Supermicro システムの構成は、NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU 向けに最適化し、認定されており、NVIDIA AI Enterprise ソフトウェアとの互換性とサポートを確保することで、プロダクショングレードの AI 開発と導入を支援します。

NVIDIA RTX PRO 6000 Blackwell Server Edition をサポートする PCIe 最適化システム

2-GPU **2U CloudDC /SYS-222C-TN**	4-GPU **2U 4U MGX /SYS-422GL-NR**	4-GPU **5U GPU SuperServer / SYS-522GA-NRT or AS -5126GS-TNRT2**
		
		Intel **AMD**
Dual Xeon 6700 CPUs, 6 PCIe 5.0 x16/x8 slots, 2 kW CRPS PSU. Up to 24 NVMe front drives in a DC-MHS chassis.	Eight GPU-ready slots; ship with four RTX PRO 6000s for a 2.4 kW draw. Plenty of NIC slots for 2 × 200 GbE or IB.	13 PCIe 5 x16 slots & 24 front U.2/U.3 NVMe bays. Taller 5U plenum + 8.6kW usablepower keeps GPUs & DPUs cool together.
2-2-3 architecture: • 2 (Dual) Intel Xeon 6700 series CPUs • 2 GPU PCIe per system (Up to 2-GPU) • 3 NIC: E/W: 2x BF3 B3140H, N/S: BF3 B3220 • Alternate NIC: CX7 2x200G	**2-4-3 architecture:** • 2 (Dual) Intel Xeon 6900 series CPUs • 4 GPU PCIe per system (Up to 8-GPU) • 3 NIC: E/W: 2x BF3 B3140H, N/S: BF3 B3220 • Alternate NIC: CX7 2x200G	**2-8-5 architecture:** • 2 (Dual) Intel Xeon 6th / AMD EPYC 4th CPUs • 8 GPU PCIe per system (Up to 8-GPU) • 5 NIC: E/W: 4x BF3 B3140H, N/S: BF3 B3220 • Alternate NIC: CX7 2x200G

2-GPU RTX PRO Server – SYS-222C-TN

4-GPU RTX PRO Server – SYS-422GL-NR

8-GPU GPU-optimized solution – AS -5126GS-TNRT2 or SYS-522GA-NRT

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

AI ワークロード向けに最適化された NVIDIA GPU

以下の NVIDIA GPU は、大規模な RAG ワークロードの高速化に最適です。NVIDIA RTX PRO™ 6000 Blackwell Server Edition、NVIDIA H200 NVL、NVIDIA HGX™ B200 および B300 の各 GPU は、メモリー、性能、相互接続機能の独自の組み合わせを提供し、企業は低レイテンシで高精度な推論を維持しながら、RAG パイプラインの規模と複雑さに合わせて GPU リソースを調整できます。

NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU

画期的な NVIDIA Blackwell アーキテクチャを基盤とする NVIDIA RTX PRO™ 6000 Blackwell Server Edition は、AI とビジュアルコンピューティング機能を強力に組み合わせ、エンタープライズ・データセンターのワークロードを高速化します。小規模モデル（70B 未満）の推論と、ファインチューニングを活用して生成 AI に加え、幅広いエンタープライズワークロードの実行を目指す企業に最適です。

- シャーシの効率的なエアフローを利用し、450 ～ 600W のパッシブヒートシンクの GPU カード複数枚を高密度に搭載することで、ラック搭載効率を向上させ、PUE を低減。
- L40/L40S/A40 GPU から、よりパワフルな PCIe Gen5 スロットに対応し、最大 8GPU 搭載でき、空冷ラックでのスケールアウトサーバー向けに最適化設計。
- エンタープライズ AI 推論、モデルのファインチューニング、HPC、仮想デスクトップ、コンテナ規模の分散グラフィックスなど、エンタープライズワークロードに幅広く対応。

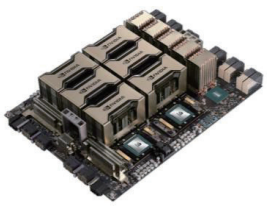
NVIDIA H200 NVL GPU

NVIDIA H200 Tensor Core GPU は、画期的なパフォーマンスとメモリー容量により、生成 AI とハイパフォーマンスコンピューティング (HPC) のワークロードを強力にサポートします。AI の基盤モデルのトレーニングや、推論用の大規模モデル (70 Billion を超えるモデル) を使用するお客様に最適です。

- 141GB で 4.8TB/ 秒の HBM3e 搭載 GPU と、900GB/ 秒の NVLink のペアにより、大規模モデル向けの 282GB 論理デバイスを形成。
- フルスケール LLM トレーニング、大規模バッチ生成 AI 推論、大規模グラフレコメンダー、メモリー容量に影響する HPC（CFD/ 気象）。
- フットプリントあたりのノード数を最大化し、トークンあたりのレイテンシと消費電力を削減しながらスループットを向上。

NVIDIA HGX プラットフォーム (HGX B300 / HGX B200)

前世代と比べて最大 30 倍の AI ファクトリー性能を備え、拡張性のある最も高性能なアクセラレート・プラットフォームである、NVIDIA Blackwell Ultra 搭載の HGX B300 システムは、最も要求の厳しい生成 AI、データ分析、HPC のワークロード向けに設計されています。集中的な AI トレーニングと推論ワークロードを抱える企業向けに設計されたこのシステムは、トレーニングと推論に利用できる最もパワフルな NVIDIA GPU です。



- ノードあたりの最大パフォーマンスを実現する 8 基の GPU と 1.8TB/ 秒の NVLink によって、大規模モデルのトレーニングや高スループットの推論などの用途に最適。
- 高速なマルチ GPU 性能のメリットを享受できる、完全な LLM トレーニング、AI 推論、大規模バッチ生成 AI、FP64/HPC 向け。
- 高 TDP（GPU あたり約 700W ～ 1200W、液冷）に対応し、優れたクロックとパフォーマンス密度を実現。

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

ステップ 2：NVIDIA AI Enterprise で AI 開発を加速

NVIDIA AI Enterprise は、NVIDIA NIM、NVIDIA NeMo™ マイクロサービスを含む、AI アプリケーションの開発、展開、拡張を加速・簡素化するソフトウェアツール、ライブラリ、フレームワークを網羅したクラウドネイティブなソフトウェアスイートです。様々な規模の組織が、広範なパートナーエコシステムを活用することで、クラウド、データセンター、エッジなど、あらゆる場所にエージェント型 AI システムを展開できます。NVIDIA AI Enterprise は、信頼性、セキュリティ、拡張性に優れた AI 運用を実現しながら、市場投入までの時間を短縮し、インフラストラクチャコストを削減します。

NVIDIA AI Enterprise 環境の概要

NVIDIA AI Enterprise は、AI アプリケーションの構築と実行のための、パフォーマンスを最適化したモジュール型環境を提供します。ライブラリ、ツール、コンテナは、モデルオーケストレーションからリアルタイム推論まで、AI パイプライン全体を高速化し、エンタープライズワークロードの迅速な展開を可能にします。クラウドネイティブな設計は、業界をリードするオーケストレーションプラットフォームと統合されており、AI アプリケーションをクラウド、オンプレミス、エッジでシームレスに実行できます。長期サポートのソフトウェアブランチ、プロアクティブなセキュリティ更新、テスト済みの展開ガイダンスにより、エンタープライズ環境にも対応します。さらに、充実したパートナーエコシステムが、ハードウェア、ソフトウェア、システム統合をサポートします。

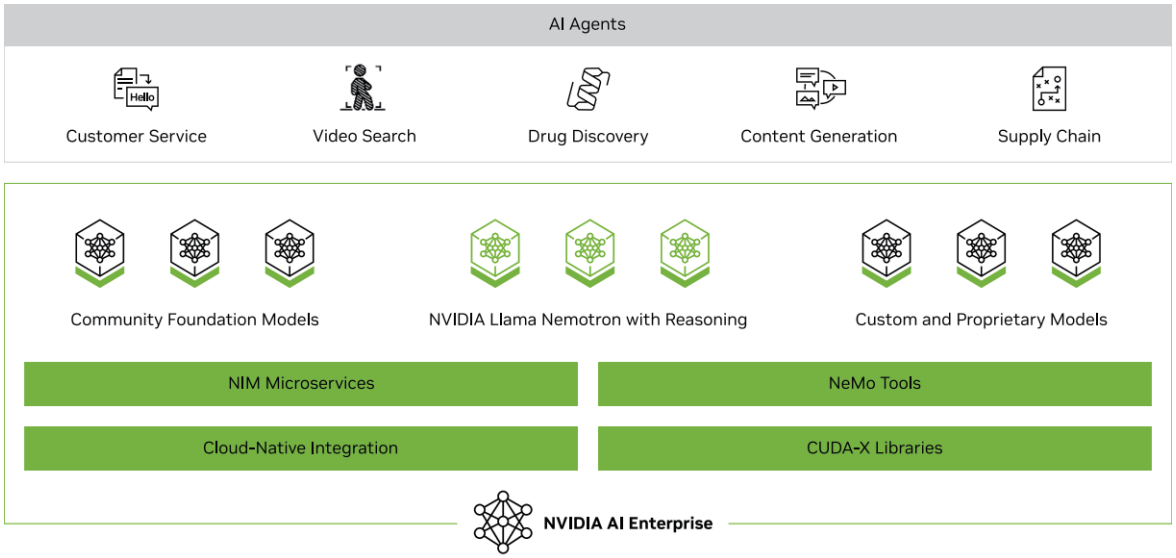
詳細については、NVIDIA AI Enterprise | Cloud-native Software Platform | NVIDIA を参照ください。または、NVIDIA AI Enterprise: Get Started With NVIDIA AI Enterprise | NVIDIA サイトから使用を開始してください。

主要な構成要素：

1. NVIDIA NIM マイクロサービス – クラウド、データセンター、ワークステーション、エッジなど、あらゆる NVIDIA アクセラレート・インフラストラクチャに最新の AI モデルを迅速に展開するための、構築済みで最適化された推論マイクロサービスを提供。
2. NVIDIA NeMo™ – 大規模言語モデル（LLM）、視覚言語モデル (VLM: Vision Language Model)、ビデオモデル、スピーチ AI など、カスタム生成 AI をあらゆる場所で開発できるエンドツーエンドのプラットフォーム。
3. NVIDIA Blueprints – 開発と導入を加速するための、検証済み AI ワークフローテンプレートのコレクション。

NVIDIA AI Enterprise の環境と、その構成要素が整ったら、次のステップでは、NVIDIA Blueprints を活用して、エンタープライズ RAG パイプラインの構築と拡張を開始します。

エージェント型 AI 向け本番環境対応ソフトウェア



© 2026 Copyright Super Micro Computer, Inc. All rights reserved

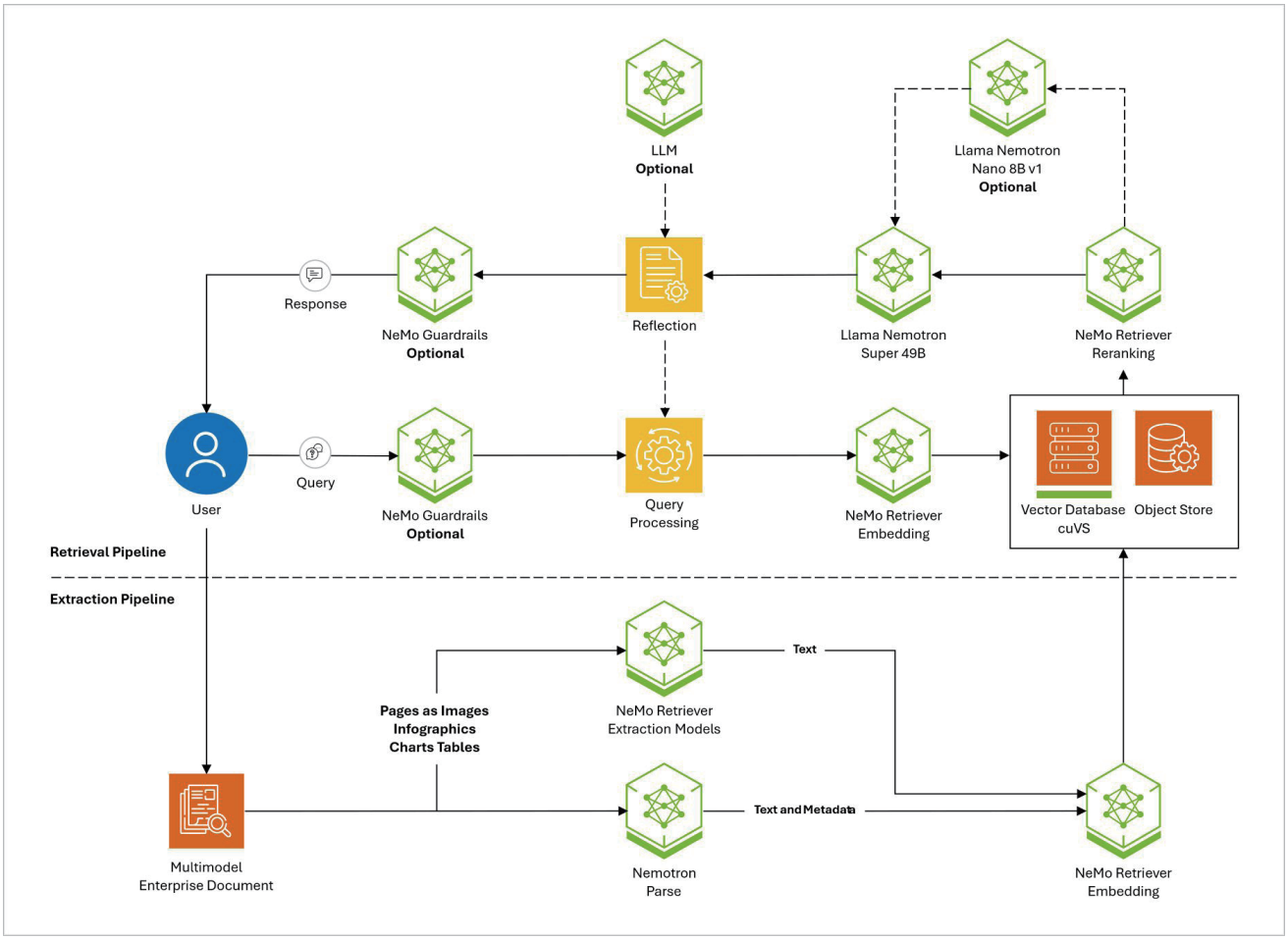
ステップ 3：NVIDIA ブループリントでエンタープライズ RAG パイプラインを構築

はじめに、Build.nvidia.com にアクセスし、NVIDIA AI Blueprint for RAG を選択します。この NVIDIA Blueprint は、NVIDIA NeMo Retriever モデルを用いて、スケーラブルでカスタマイズ可能なデータ抽出・取得パイプラインを構築するための基礎的な出発点を開発者に提供します。NVIDIA Blueprint を使用することで、コンテキストアウェアな応答を提供するための、テキスト、表、グラフ、数百万件もの PDF など、広範なマルチモーダルエンタープライズデータに接続するなど、LLM を広範なマルチモーダルエンタープライズデータに接続することができます。企業は、マルチモーダル PDF データ抽出を 15 倍高速化し、誤回答を 50% 削減することで、実用的な洞察を獲得し、生産性向上を大規模に実現できます。

RAG 向け NVIDIA AI ブループリントは、事前トレーニング済みの LLM と、ターゲットを絞ったデータ取得を組み合わせた、エンタープライズ規模の AI ソリューションを構築するための、本番環境対応のワークフローを提供します。NVIDIA NeMo Retriever と Llama Nemotron モデルを搭載し、高い精度、強力な推論能力、エンタープライズ規模のスループットを実現し、プロトタイプから本番環境への移行を、数ヶ月ではなく数週間で実現します。高度な取得、再ランク付け、リフレクションの技術により、幻覚現象を軽減し、結果が社内データやポリシーと整合していることを保証します。ブループリントには、機密情報を保護するためのガバナンス、可観測性、安全性の機能も含まれており、GPU アクセラレーションによって、大規模環境でも信頼性と回復力に優れたパフォーマンスを実現します。柔軟なプラグインとカスタマイズ機能により、エンタープライズ検索、ナレッジアシスタント、ジェネレーティブ・コパイロット、垂直 AI ワークフローなど、スタンドアロンまたはより高度なエージェント型アプリケーションに統合したソリューションに適応させることができます。

2 What Is Agentic AI? | NVIDIA Blog

アーキテクチャとワークフローを表す図



© 2026 Copyright Super Micro Computer, Inc. All rights reserved

このモジュール設計により、効率的なクエリ処理、正確な情報取得、そして容易なカスタマイズが実現します。

プロセス、主要機能、システム要件をより深く理解するために、GitHub documentation を確認することをお勧めします。そこから、ドキュメントに記載されている手順に従ってオンプレミス、または、クラウドで展開を開始できます。

まとめ

適切な基盤が整えば、エンタープライズ対応の RAG ソリューションを活用した AI 導入を加速させることができます。企業における AI 導入の課題、専用 AI ファクトリー・インフラストラクチャの威力、そして検証済みの RAG ブループリントを基盤とする Supermicro と NVIDIA の統合プラットフォームが、高度な AI 導入をよりシンプル、迅速、そしてスケーラブルに実現する方法について解説しました。Supermicro と NVIDIA は AI 変革をリードし、企業が AI の潜在能力を現実世界への影響へと変えるために必要なツールとプラットフォームを提供します。

モデルの種類やサイズの最適化、検索ライブラリの拡張、同時ユーザーの管理、入出力コンテキストウィンドウの調整など、Supermicro と NVIDIA、そして信頼できるパートナーが、お客様固有の要件に合わせてソリューションをカスタマイズするための専門知識とインフラストラクチャを提供します。この協業アプローチにより、お客様のビジネス目標に沿った、スムーズで効果の高い導入を実現できます。

今すぐ RAG の展開を始めましょう

- **Supermicro と RAG 準備コンサルテーションをご予約いただき、お客様固有のワークロードに合わせて、インフラストラクチャ、GPU、ソフトウェアをカスタマイズ。**
- **Build.nvidia.com で AI ブループリントを探し、実証済みのアーキテクチャと導入戦略をご確認。**
- **RAG 向け NVIDIA AI ブループリントにアクセスし、ただちに、検索拡張生成パイプラインの構築を開始。**

お問い合わせ：スーパーマイクロ株式会社 〒 150-0031 東京都渋谷区桜丘町 20-1 渋谷インフォスタワー 21 階
電話：03-5728-5196 FAX：03-5728-5197 Email：Sales_Inquiry_JP@Supermicro.com

© 2026 Copyright Super Micro Computer, Inc. All rights reserved