NVIDIA NIMで実現する生成AIの推論パフォーマンス最適化

Eriko Tsuda, Software BizDev | NVIDIA Japan | Feb 2025

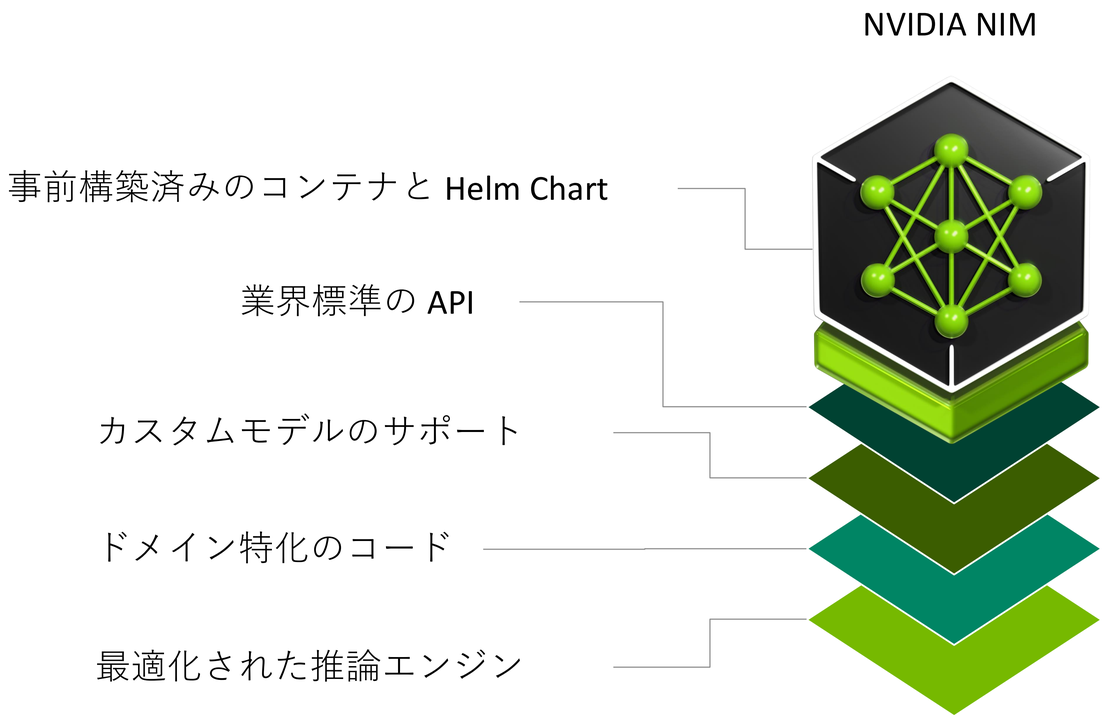
Accelerated Computing Platform Company

- 1993 年創業
- 創業者及び CEO　ジェンスン フアン
- 従業員 32,000 人
- 2024会計年度売上高 609億 ドル

NVIDIA NIM

NVIDIA NIM - 最適化された推論マイクロサービス

生成 AI のための高速化されたランタイム



AIアプリケーションとデータを
セキュリティと制御を維持しどこにでもデプロイ

ビルド済みかつ継続的にメンテナンスされる
マイクロサービスにより**市場投入までの時間を短縮**

最新の AI モデル、標準 API、エンタープライズグレードのツールにより**開発者に力を与える**

最適化されたスループットとレイテンシにより
トークン生成速度とレスポンス速度を最大化

企業独自のデータを用いたモデルチューニングにより**精度を向上**

プロダクションデプロイのための安定した API、セキュリティパッチ、QA、エンタープライズサポート

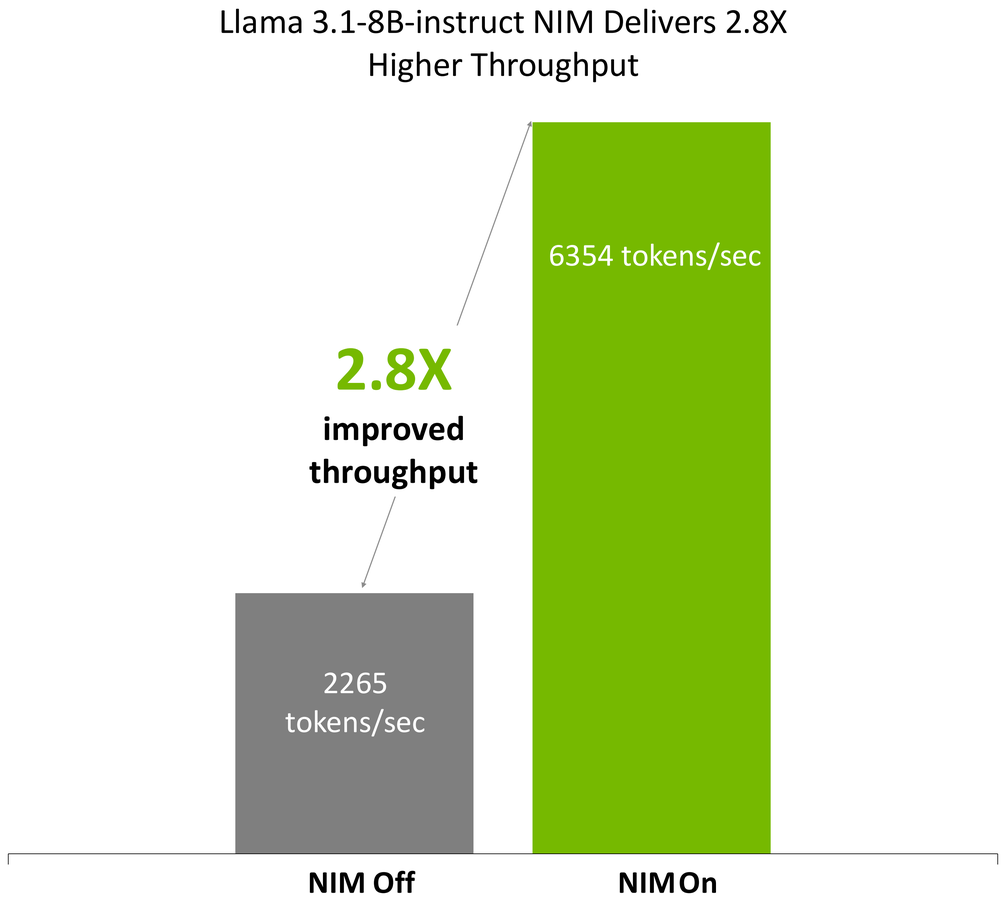
NVIDIA NIM はAI推論のための最速の道

最適化された高速なモデルデプロイのためのエンジニアリングコストを削減

	NVIDIA NIM	Do It Yourself
デプロイまでの時間	5分	1週間以上
API標準	業界標準API OpenAI for LLMs, Google Translate for Speech	各ドメインとモデルのAPIレイヤーを 業界標準仕様に従って実装
最適化されたエンジン	事前ビルド済みのNVIDIAおよびコミュニティモデル MISTRAL AI_ Meta starcoder NVIDIA Nemotron	ユーザーが自身でエンジンをビルドし、 ハードウェア特有の制約に応じてカスタマイズ
前処理/後処理	前後処理(tokenization)のために最適化された 事前構築済みのパイプライン	カスタムロジックの実装
モデルサーバーの デプロイ	自動	手動でセットアップ
カスタマイズ	LoRA, SFTをサポート	カスタムロジックの実装
コンテナの検証	広範なワークロードに特化したマトリクス検証によるQA	事前検証なし
サポート	NVIDIA AI Enterpriseによる提供 CVEスキャン/パッチ適用、技術サポート	サポートなし

NVIDIA NIMの圧倒的なスループット

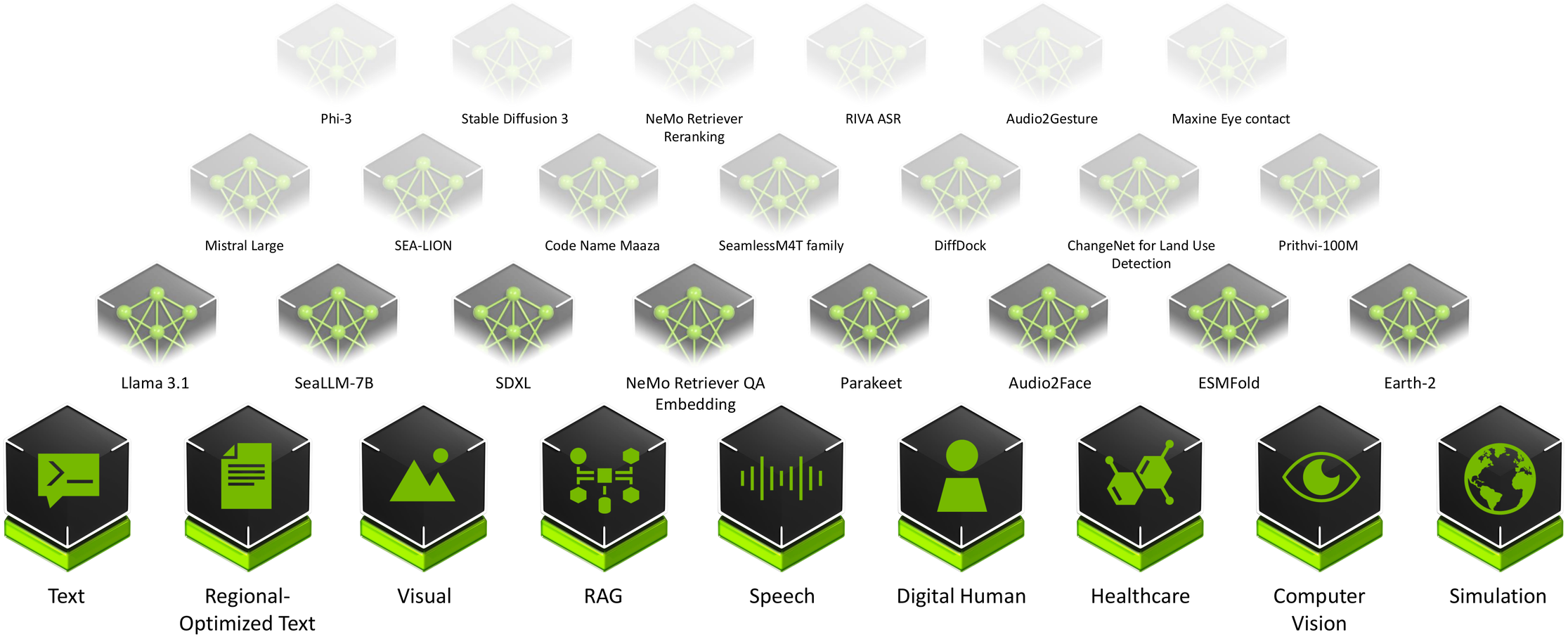
最速スループットでソリューション全体のコストを削減



Llama 3.1-8B-instruct, input token length: 1,000, output token length: 1,000. Concurrent client requests: 200. 1xH100 SXM NVLink. NIM Off: FP8, TTFT: NIM On: FP8. ~.4s, ITL: ~31ms.

生成AIのための推論マイクロサービス

NIMは、クラウド、データセンター、ワークステーションでAIモデルを展開する最速の方法

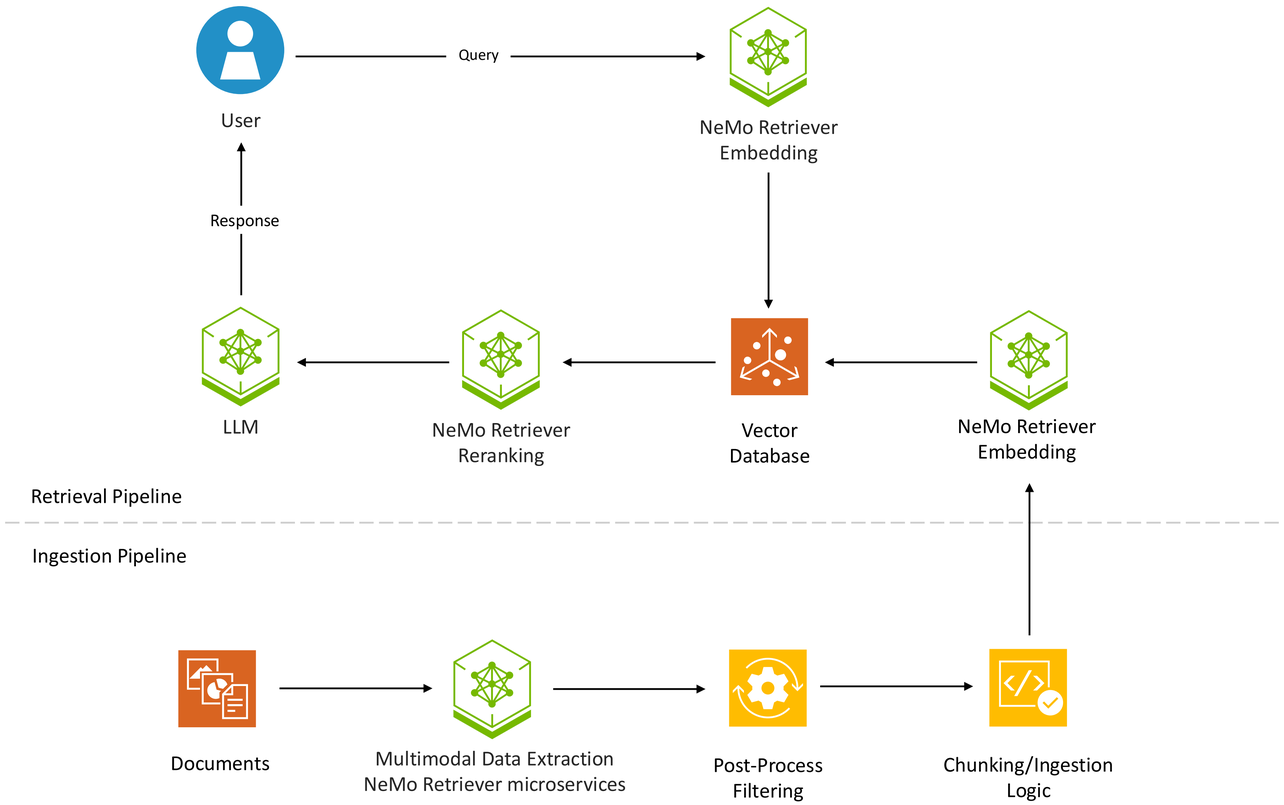


[NVIDIA API Catalog]

NVIDIA NIM Retrieval Models for RAG

NeMo Retriever が RAG アプリケーションを高速化

NVIDIA Open, Commercial microservices Power Enterprise RAG Pipelines



- NVIDIA NIMを基盤とした最適化された推論エンジン
- より良い精度のためにFine-tuneされた最先端のカスタマイズ可能なモデル
- 柔軟でモジュール型の展開
- 高速化されたベクター検索
- 本番環境に対応

NeMo Retriever: 情報検索のためのマイクロサービス群

Available on build.nvidia.com

nv-embedqa-e5-v5
テキスト質問応答のための埋め込みモデル

nv-embedqa-mistral7b-v2
多言語テキスト埋め込みモデル

snowflake-arctic-embed-l
最適化されたコミュニティ・モデル

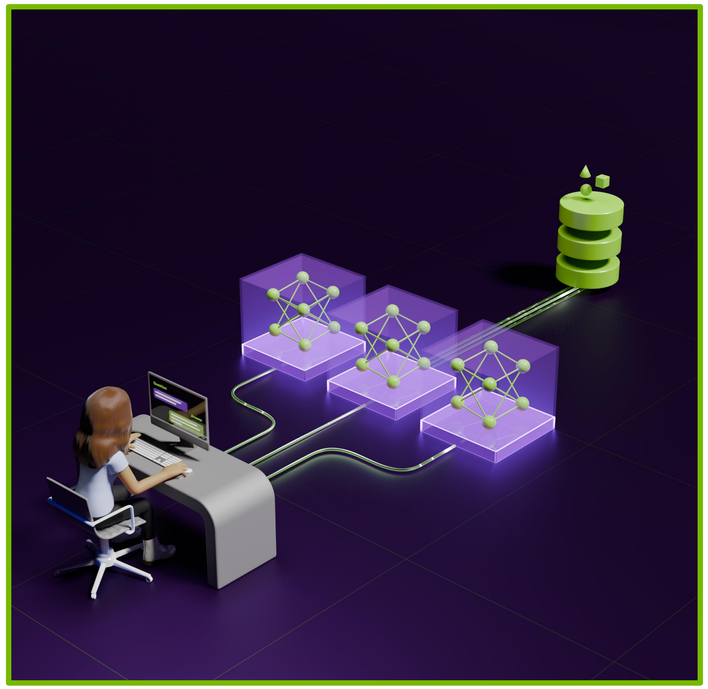
nv-rerankqa-mistral4b-v3
質問回答のためのテキスト リランキングモデル

New!!

llama-3.2-nv-embedqa-1-v1
世界最高水準の多言語・異言語対応した
質問応答のための埋め込みモデル

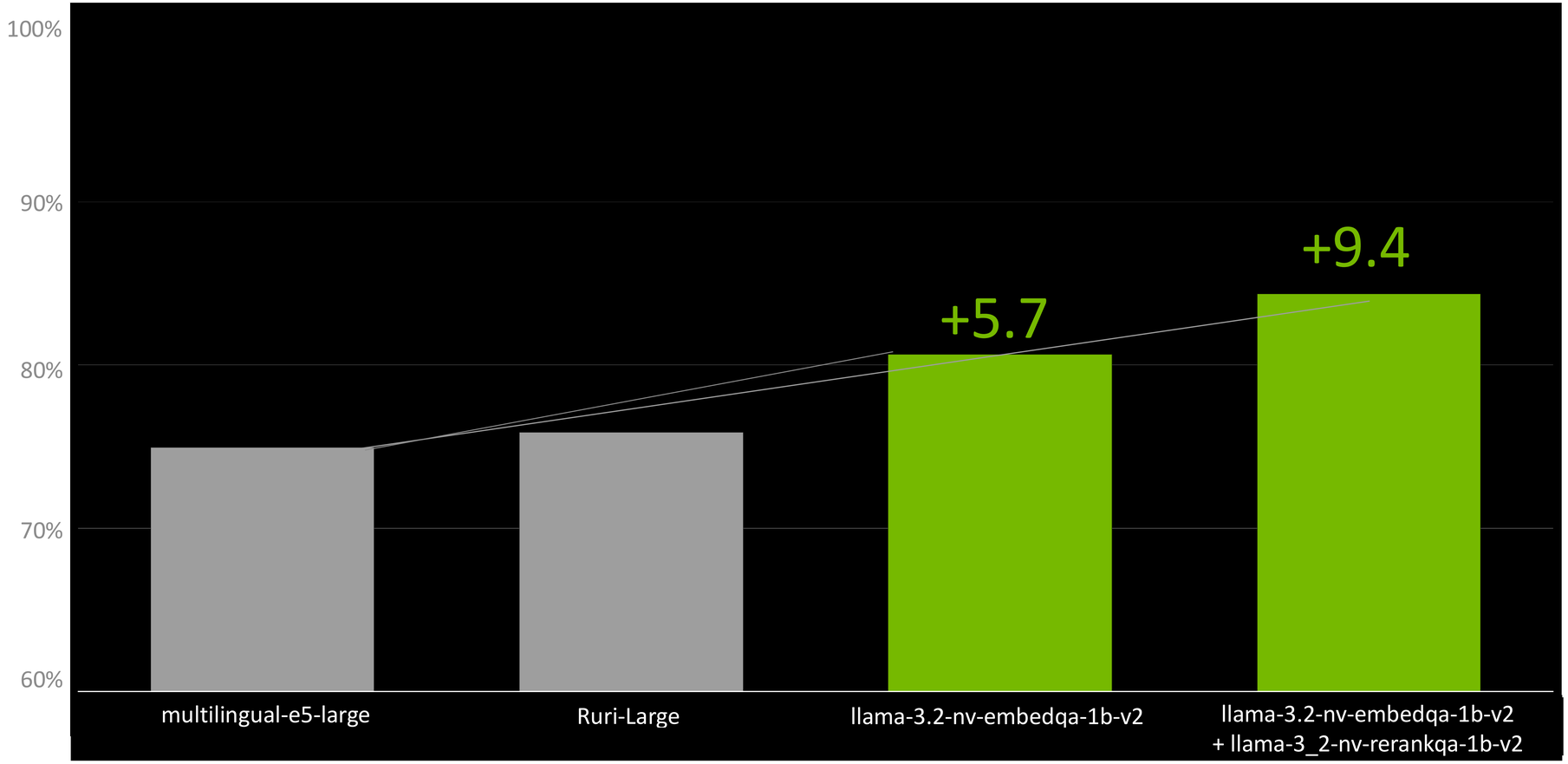
New!!

llama-3.2-nv-rerankqa-1b-v1
複数のソースと言語で検索結果を効率的に絞り込む為の
リランキングモデル



NVIDIAの多言語対応の埋め込みモデルとリランキングモデルにより最高水準の日本語Retrievalスコアを達成

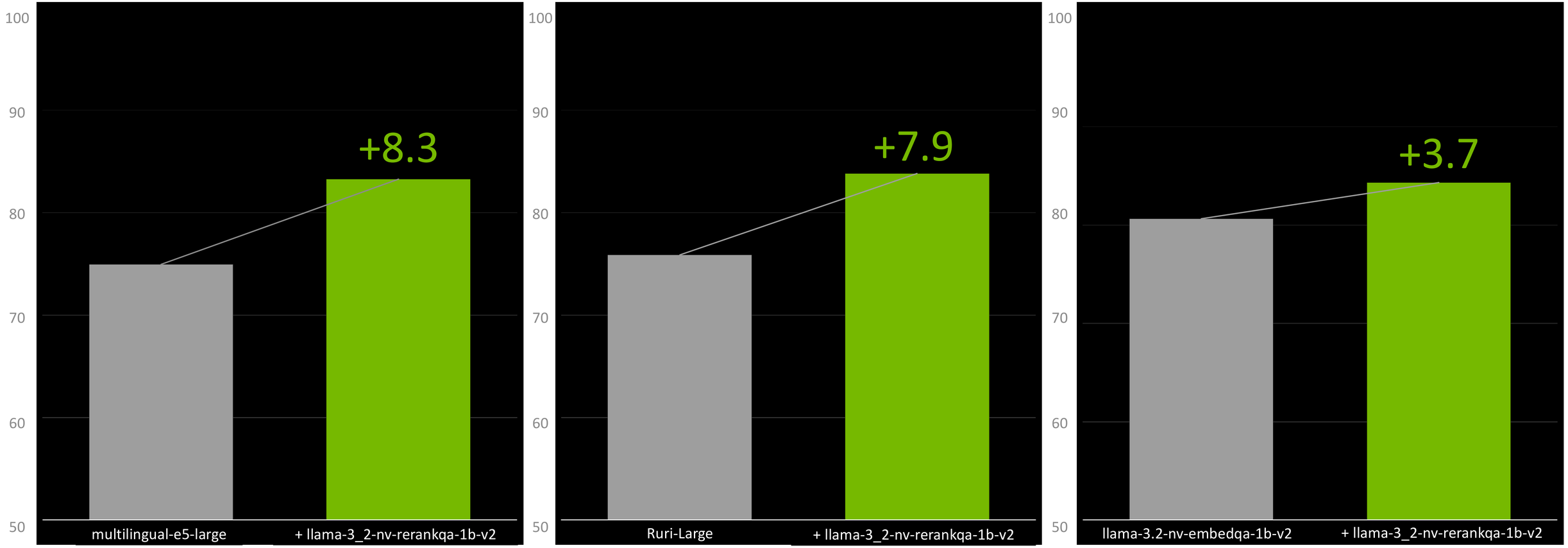
Embedding Model Accuracy on JMTEB leaderboard Retrieval Task (nDCG@10)



Accuracy Avg. of 6 datasets (JaGovFaqs, JAQKET, Mr.TyDi-Ja, NLP Journal Abs-Intro, NLP Journal Title-Abs, NLP Journal Title-Intro) on JMTEB leaderboard with an nDCG@10
Gray bars are generated by OSS SOTA embedded models,,the far-right bar is generated from a multi-stage llama-3.2-nv-embedqa-1b-v2 + llama-3.2-nv-rerankqa-1b-v2 retrieval system.

NVIDIAのリランキングモデルの追加によりJMTEBのスコアが最大8.3ポイント上昇

Embedding Model Accuracy on JMTEB leaderboard Retrieval Task (nDCG@10)



multilingual-e5-large　　Ruri-Large　　llama-3.2-nv-embedqa-1b-v2

Accuracy Avg. of 6 datasets (JaGovFaqs, JAQKET, Mr.TyDi-Ja, NLP Journal Abs-Intro, NLP Journal Title-Abs, NLP Journal Title-Intro) on JMTEB leaderboard with an nDCG@10
Gray bars are generated by OSS SOTA embedded models,,the far-right bar is generated from a multi-stage llama-3.2-nv-embedqa-1b-v2 + llama-3.2-nv-rerankqa-1b-v2 retrieval system.

Summary

JMTEBでの評価結果

- Llama-3.2-nv-embedqa-1b-v2 (埋め込みモデル): API Catalog Link
  - 他のオープンソースの埋め込みモデルと比較しても高い精度
  - 既存の埋め込みモデルから置き換えるだけで、大幅に精度向上が期待できる
  - リランキングモデル ”llama-3_2-nv-rerankqa-1b-v2”との組み合わせで更なる精度向上が見込める

- Llama-3.2-nv-rerankqa-1b-v2 (リランキングモデル): API Catalog Link
  - 埋め込みモデル “Llama-3.2-nv-embedqa-1b-v2”との組み合わせで日本語検索精度において最高水準の精度を達成
  - 他の埋め込みモデルのRAG Pipelineに、このリランキングモデルを追加するだけでも検索精度の更なる向上が見込める

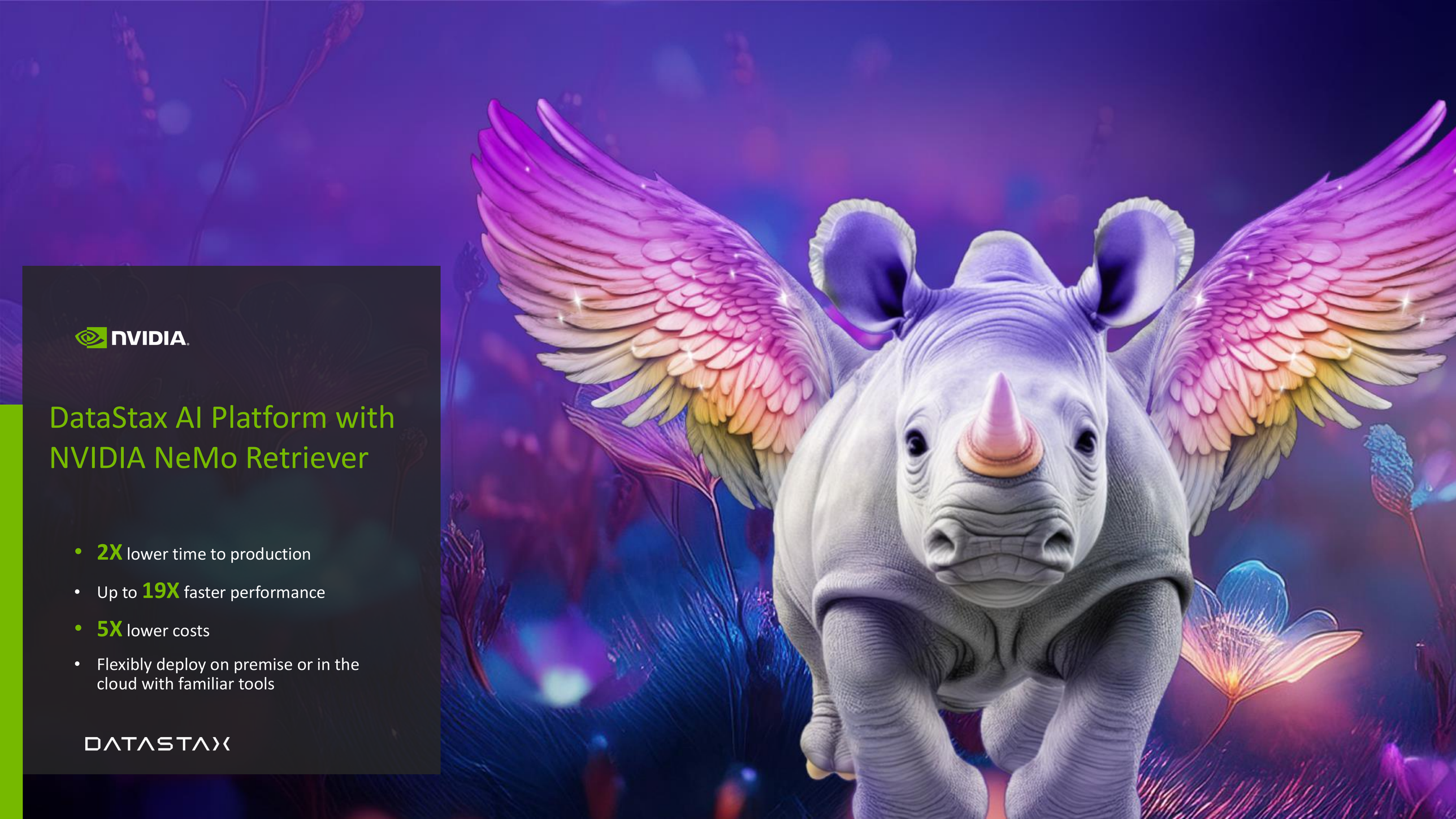
Case Study: Cadence Design Systems

3.3x fewer incorrect answers retrieving from technical documentation

Recall	Top 1	Top 3	Top 5	Top 10
Reference Pipeline	36%	52%	57%	64%
NeMo Retriever Hybrid Search	57%	70%	77%	80%
NeMo Retriever Hybrid Search + Reranker	69%	81%	86%	89%
Improvement Factor	**2x**	**2.5x**	**3x**	**3.3x**

cadence®

NVIDIA



NVIDIA

DataStax AI Platform with NVIDIA NeMo Retriever

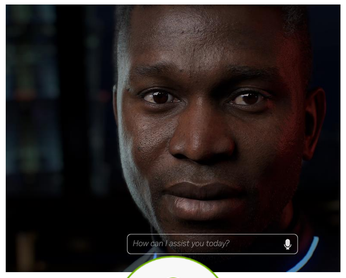
- **2X** lower time to production
- Up to **19X** faster performance
- **5X** lower costs
- Flexibly deploy on premise or in the cloud with familiar tools

DATASTAX

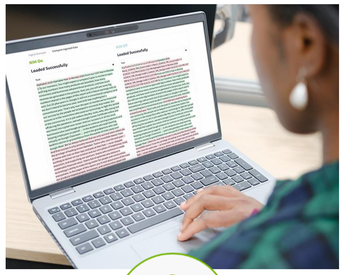
NVIDIA Blueprints include NeMo Retriever

Available on build.nvidia.com

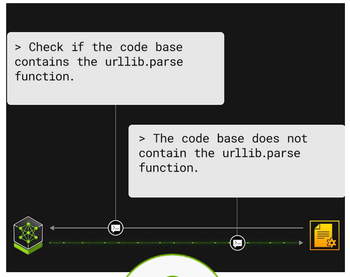
カスタマーサービス向け
デジタルヒューマン



企業でのRAG構築向け
マルチモーダルPDF抽出

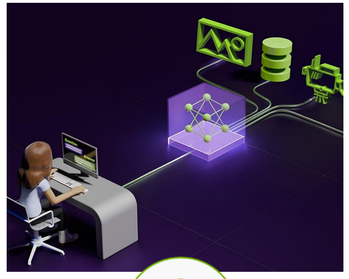


コンテナセキュリティのための
全弱性分析



カスタマーサービス向け
AIバーチャルアシスタント

ビデオ検索と要約のための
ビジュアルAIエージェント



サンプルアプリ

Interactive experience that can be easily replicated

サンプルデータ

Public data for workflow testing

リファレンスコード

Leverage proven pre-trained models

アーキテクチャ

Reference architecture including API definitions, NIM, and more

カスタマイズツール

Customize and evaluate models

オーケストレーション

Deploy and manage workflow microservices

NIM Access and Licensing

build.nvidia.com または ai.nvidia.com で無料でお試し

- NVIDIA NIMは API Catalog 経由で利用可能
  - まずは API Catalog に掲載のいずれかのエンドポイントへアクセス (ex: https://build.nvidia.com/meta/llama-3_1-70b-instruct)
  - “Build with this NIM” をクリック
- ライセンシング
  - NIM は NVIDIA AI Enterprise にてライセンス
    - NVIDIA AI Enterprise の90日間無償ライセンスで API Catalog から利用可能 (最大 5000 クレジット – 1クレジット = 1 API call)
    - NIM のプロダクション展開については、年間サブスクリプションの購入が必要 (1ライセンス / 1 GPU)
  - NVIDIA Developer Program のメンバーは、研究/開発/テスト目的であればダウンロード可能なNIMを無料で使用可能