

NVIDIA releases Nemotron 3.5 retrieval models for agent systems

Source copy kept by wheresthe.ai for the story "NVIDIA releases Nemotron 3.5 retrieval models for agent systems" (<https://wheresthe.ai/d/intl-nvidia-releases-nemotron-3-5-retrieval-models-for-agent-2026-10/>).

Source: <https://huggingface.co/blog/nvidia/nemotron-retriever-world-class-agentic-retrieval>

Text saved 21 Oct 2026 (the page as published).

SHA-256 of the saved text: 206fc80a5e134b46ffa2ab04e9f31c5be19ba2bbaf74c07a39206fdeac1b804b

Back to Articles a]:hidden">

NVIDIA Nemotron 3.5 Embed, Rerank, and Parse: Advancing Agents with World-Class Retrieval Models

Community Article Published October 21, 2026

Upvote 9

+3

Yauhen Babakhin ybabakhin Follow

nvidia

Ryan Chesler emelryan Follow

nvidia

Jiarui Cai jiaruic Follow

nvidia

Sean Sodha ssodha-nv Follow

nvidia

Isabel Hulseman ihulseman0220 Follow

nvidia

:last-child]:mb-0">

An agent's answers depend on the evidence it can retrieve. When that evidence is incomplete or irrelevant, the agent will repeat searches and inspect additional documents, consuming more inference tokens before completing the task. Better retrieval can help agents reach grounded answers with fewer wasted tokens. At enterprise scale, that waste leads to massive inefficiencies. Agents need the right retrieval models and systems to reach insights faster.

Today, we are expanding the Nemotron Retriever collection with new open, commercially available models for embedding, reranking, and document extraction. The new models include Nemotron 3.5 Embed 8B BF16 , 1B BF16 , 1B NVFP4 , Nemotron 3.5 Rerank 8B BF16 , and Nemotron 3.5 Parse at 631M parameters.

Together, these state-of-the-art models bring multimodal retrieval, relevance ranking, and document processing to enterprise agent workflows.

Better Retrieval for More Efficient Agents

Retrieval quality affects both the answers an agent produces and the LLM inference cost. Relevant evidence can help the agent answer a question or better decide its next step, and reranking improves the quality of evidence that enters its context.

Consider an agent asked to identify products at risk of stockouts next quarter and explain why. It must combine demand forecasts in charts, inventory levels in tables, and supplier lead times in reports, then reason about whether replenishment will arrive before inventory runs out. Completing the task requires finding the relevant evidence across documents and using it together in the analysis.

To understand how retrieval affects the agent's ability to combine evidence and complete a task, we evaluated

Nemotron 3.5 Embed and Rerank across ViDoRe v3 , BRIGHT , and BrowseComp-Plus . ViDoRe v3 uses combined text and image inputs, while BRIGHT and BrowseComp-Plus use text-only inputs. In this pipeline, an agent queries the corpus using a search tool, inspects the evidence, and refines its next search attempt to find the most relevant information to a given query. The experiments below vary how the search tool is implemented: embedding only, embedding plus reranking, and with different models for each. The LLM powering the agent, tools, task set, corpus, and execution policy remain fixed while the retrieval configuration changes. We measure the agent's ability to find the relevant information for each task and total agent inference tokens per task, including input and output tokens across all turns.

Figure 1. End-to-end task success versus total agent inference tokens per task across retrieval configurations.

Lower agent inference cost: Adding Nemotron 3.5 Rerank reduces agentic token cost per query by approximately 14% with Nemotron 3.5 Embed 8B, and 13% with Nemotron 3.5 Embed 1B, while improving retrieval accuracy in both configurations.

Better evidence within the same agent budget: With Nemotron 3.5 Embed 8B as the retriever, Nemotron 3.5 Rerank delivered approximately 0.9 points higher nDCG@10 than Qwen3-VL-Reranker-8B at a similar agentic token cost.

Nemotron 3.5 Embed and Rerank

Behind these agent successes are Nemotron 3.5 Embed and Nemotron 3.5 Rerank, which retrieve evidence across text and images and prioritize it for the agent's context. Nemotron 3.5 Embed builds on the previous generation's Nemotron 3 Embed text embedding models to retrieve evidence from document images, including tables, charts, infographics, datasheets, and schematics. This allows applications to search information conveyed by visual content alongside document text.

The four models provide different options for retrieval and second-stage ranking:

Model Role Deployment focus

Nemotron-3.5-Embed-8B-BF16 Retrieves evidence across text and images, with strong multilingual retrieval capabilities. Quality-focused search over text and visually rich documents

Nemotron-3.5-Embed-1B-BF16 Distills the 8B model's multimodal retrieval capabilities into a compact model. Multimodal indexing and query serving with a smaller model footprint

Nemotron-3.5-Embed-1B-NVFP4 Applies NVFP4 quantization to both text and image encoders for efficient multimodal inference. Large-scale document indexing and high-throughput retrieval on supported hardware

Nemotron-3.5-Rerank-8B-BF16 Jointly evaluates query–document pairs to refine relevance ranking across text and image candidates. Selecting relevant evidence from retrieved candidates for RAG and agentic workflows

Table 1. Nemotron 3.5 Embed and Nemotron 3.5 Rerank model capabilities and use cases across various sizes and variants.

Key features include:

Text and image retrieval: Search document text and images containing tables, charts, infographics, datasheets, and schematics.

Multilingual support: Text retrieval was evaluated across 34 languages, while multilingual visual retrieval was evaluated on MIRACL Vision (18 languages) and KoViDoRe. The model may support additional languages beyond those evaluated.

Multimodal reranking: An 8B cross-encoder scores query–document pairs across text and image inputs.

BF16 and NVFP4 deployment: The 1B embedding model is available in BF16 and NVFP4, with quantization applied to both text and image encoders.

Domain adaptation: Nemotron embedding fine-tuning recipes support custom document collections, with a 18% and 3% relative nDCG@10 gain for Nemotron-3.5-Embed-1B-BF16 and Nemotron-3.5-Embed-8B-BF16 respectively in our chart and infographic experiments.

Leading Multimodal Dense Retrieval on ViDoRe v3

On ViDoRe v3.1 , Nemotron 3.5 Embed 8B BF16 achieves 61.17 nDCG@10, ranking first among dense retrieval models. This is an 8.94-point improvement over the 52.23 score achieved by the Llama Nemotron Embed VL 1B model. The 1B model brings most of this visual retrieval quality into a smaller model. Distilled from the 8B teacher, Nemotron 3.5 Embed 1B BF16 scores 58.38 nDCG@10, retaining 95.4% of the teacher's score.

Figure 2. Nemotron 3.5 Embed 8B BF16 ranked first among dense retrieval models on the ViDoRe v3.1 leaderboard as of 10/20/2026.

We also compare the models with Qwen3-VL-Embedding, EmbeddingGemma-2, and the previous-generation NVIDIA model across ViDoRe v3 , Korean ViDoRe and MIRACL Vision . The chart below reports retrieval accuracy for image-only and combined image-and-text inputs.

Figure 3. Text and visual retrieval performance of Nemotron 3.5 Embed compared with prior-generation models and selected baselines.

The multilingual results show how the models retrieve evidence across different language settings. On MIRACL Vision , the 8B model achieves 72.6 nDCG@10, exceeding Qwen3-VL-Embedding-8B by 4.3 points. We also evaluate the models on KoViDoRe ; both Nemotron 3.5 Embed 8B and 1B outperform the Qwen baselines when using combined image and text inputs.

Reranking the Evidence for Agent Tasks

Embedding models make large collections searchable by representing content as vectors. After retrieval returns a candidate set, Nemotron 3.5 Rerank 8B BF16 evaluates each candidate against the query to refine the ordering across text and image inputs. By distinguishing broadly related content from evidence that directly answers the question, the reranker helps agents use their context budget more effectively and focus on information needed to complete the task. For agentic retrieval systems, surfacing the strongest evidence earlier can reduce distracting context and help agents produce more grounded, efficient responses.

In collaboration with the MTEB team, we are launching the ViDoRe v3.1 reranking leaderboard to benchmark rerankers on document retrieval across image-only, text-only, and combined image-and-text inputs. Nemotron 3.5 Rerank 8B BF16 achieves 60.49 nDCG@10 with combined image-and-text inputs compared with Qwen-3-VL-Reranker-8B's 59.17.

Figure 4. Nemotron-3.5-Rerank-8B-BF16 ranked first on ViDoRe v3.1 Reranking leaderboard as of 10/20/2026.

Figure 5. Text and visual reranker performance of Nemotron 3.5 Rerank 8B compared with Qwen 3 VL Reranker 8B. Both models reranked the top 50 candidates retrieved by the Nemotron 3.5 Embed 1B model.

Scaling Multimodal Retrieval with NVFP4

Retrieval quality is only part of the equation for production agentic systems. As document collections and query volumes grow, teams must balance accuracy, throughput, and infrastructure requirements.

Nemotron 3.5 Embed 1B NVFP4 provides an optimized deployment option for high-throughput multimodal retrieval. On NVIDIA GB200, it delivers 27.1% higher throughput compared with the 1B BF16 model, increasing from 101.3 to 128.8 documents per second while retaining 99.85% of its retrieval accuracy on ViDoRe v3 with combined image and text inputs. As Figure 6 shows, the NVFP4 model offers a strong accuracy-throughput tradeoff for teams scaling retrieval across production workloads.

This deployment flexibility complements domain-specific customization: teams can select the data type of model weights that fits their infrastructure, then adapt the embedding and reranking models to their own terminology, documents, and retrieval tasks.

Figure 6. ViDoRe V3 retrieval accuracy versus serving efficiency, comparing Nemotron 3.5 Embed with selected smaller open embedding baselines.

High-Performance Serving with NVIDIA NIM

NVIDIA NIM provides a high-performance deployment path for Nemotron 3.5 Embed. In testing on an NVIDIA RTX PRO 6000 Blackwell Server Edition, NIM delivered 6,031 query requests per second—approximately 2.9× the throughput of a comparable vLLM configuration at 2,099 requests per second—while maintaining comparable passage-processing throughput. This optimized serving stack helps enterprises scale retrieval workloads with production-ready deployment and support.

Figure 7. Nemotron 3.5 Embed NIM serving performance compared with the vLLM checkpoint on NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs.

Adapting Retrieval to Domain-Specific Tasks

Enterprise agents often search documents containing specialized terminology and company-specific information that general-purpose retrieval models may not fully capture. Fine-tuning allows teams to adapt Nemotron 3.5 Embed and Rerank using their own documents, queries, and relevance labels to their specific domains and use cases.

The Nemotron multimodal embedding fine-tuning recipe combines synthetic data generation, hard-negative mining, fine-tuning with NVIDIA NeMo AutoModel , and evaluation in one workflow. Teams can generate queries and relevance labels from their own documents to build training data for their retrieval tasks.

We evaluated the recipe on 8,202 chart and infographic images using 8,591 synthetic queries generated with Qwen3.8-Flash-Next, focusing on enumerative and open-ended queries with multiple relevant images. Fine-tuning the Nemotron-3.5-Embed-1B-BF16 on 6,872 queries for one epoch increased nDCG@10 from 37.34 to 44.12, a 18% relative gain over the base model, on 1,719 held-out synthetic queries. Using the same synthetic queries set, fine-tuning Nemotron-3.5-Embed-8B-BF16 for epoch increased nDCG@10 from 41.44 to 42.75, a 3% relative gain.

How We Built the Embed and Rerank Models

Nemotron 3.5 Embed 8B builds on Ministral-3-8B-Instruct , with its language component initialized from Nemotron-3-Embed-8B-BF16 . The model is trained on image and text corpora using contrastive learning to produce representations for multimodal retrieval.

The Embed 1B models are distilled from Nemotron 3.5 Embed 8B, with their language component initialized from Nemotron-3-Embed-1B-BF16 . Distillation combines cosine and mean squared error (MSE) losses to align the student's embeddings with the teacher's.

The language component remains frozen throughout embedding-model training. Text-only embeddings are therefore expected to remain unchanged from the corresponding Nemotron 3 Embed models, allowing existing text embeddings and vector indexes to be reused.

Nemotron 3.5 Rerank uses the same backbone architecture as Nemotron-3.5-Embed-8B in a cross-encoder configuration. While the embedding model encodes queries and documents separately and scores their similarity, the reranker jointly encodes each query–document pair. A final multilayer perceptron (MLP) scoring head produces a scalar relevance logit.

All models are trained on a mixture of publicly available and synthetic multilingual, multimodal data, and their final released weights are produced through checkpoint ensembling. Further details on the training data are available in each model's Hugging Face model card.

Nemotron 3.5 Parse

Enterprise knowledge often arrives in PDFs whose embedded text captures only part of what is on the page. Extracting that text can lose reading order, separate table values from their headers, or omit information conveyed in charts. Scanned documents and image-based PDFs may contain no usable text layer at all. Before this information can support search, retrieval, or automated document workflows, the pixels need to become structured content.

Nemotron 3.5 Parse converts document images into machine-readable text while preserving the structure that gives the content meaning. It identifies document elements, their locations, and reading order, and extracts tables and charts into structured representations. The resulting Markdown, HTML tables, and structured blocks give developers content that is better suited to indexing, search, and downstream processing. Applications can retrieve a passage with its surrounding context, process a table as rows and columns, and link extracted evidence back to its location on the original page.

At 631 million parameters, Nemotron 3.5 Parse brings this capability to document pipelines at scale, complementing Nemotron 3.5 Embed and Rerank with structured information for retrieval and agent workflows.

Leading Accuracy with Faster Document Processing

Figure 8. Nemotron 3.5 Parse ranked first on ParseBench leaderboard among models with fewer than 1 billion parameters as of 10/20/2026.

Nemotron 3.5 Parse ranks #1 on ParseBench among models with fewer than 1 billion parameters, achieving a score of 75.20 while processing 10.5 pages per second on one NVIDIA H100. It delivers best-in-class speed with well optimized parameter allocation, multi-token prediction, speculative decoding and flexible native-resolution handling. ParseBench tests table and chart extraction, content faithfulness, semantic formatting, and visual grounding.

Developers can scale processing toward speed or accuracy using the same model. Single-pass inference extracts each page once for the highest throughput. Double-pass inference revisits tables and charts identified in the initial page parse. Triple-pass inference adds a detection pass to find additional regions for refinement. This approach concentrates additional processing on document regions that benefit from a closer look.

Processing mode ParseBench score ↑ Pages per second on one H100 ↑

Single pass 72.74 16.1

Double pass 74.92 11.2

Triple pass 75.20 10.5

Table 2. Model accuracy is able to scale with additional model passes at a small cost to throughput. Reprocessing focused crops of tables and charts leads to increased accuracy.

All three configurations are Pareto optimal within this comparison: no other compared configuration achieves both higher accuracy and higher throughput. Teams can choose the operating point that fits their workload, from rapid ingestion of large document collections to more detailed extraction of complex pages.

Figure 9. Nemotron Parse 3.5 achieves pareto optimal performance on ParseBench leaderboard.

How We Built Nemotron 3.5 Parse

Nemotron 3.5 Parse is trained on a diverse mixture of enterprise documents, scientific papers, forms, tables, charts, and multilingual text. A strong synthetic data engine expands the range of layouts, structures, and languages the model sees, complementing automatically labeled documents and human-labeled data. Synthetic multilingual data helps build broad OCR capability, with strong performance in evaluations spanning 163 language and script combinations. This diversity prepares the model for the varied content found in enterprise document collections. Model card and multilingual coverage .

We further improve robustness and reliability through Group Relative Policy Optimization (GRPO), using custom reward functions to reinforce correct document structure and penalize failures. This post-training reduces repetitive generation loops and unclosed blocks, improves table consistency, and strengthens adherence to chart schemas, making extracted information more reliable for downstream document workflows.

Thank you to LightOn for raising the open-source bar with their models and technical blog . The shared techniques were pivotal in improving our model.

Early Enterprise Partner Evaluations

Elastic customers can access NVIDIA Nemotron Retriever embedding and reranking models through Elasticsearch's Inference API using NVIDIA inference endpoints. Nemotron 3 will be available via Elastic Inference Service (EIS) soon, while support for Nemotron models on-prem deployment is on the roadmap.

Vectorize evaluated NVIDIA Nemotron 3.5 Embed using Hindsight Agent Memory , and it ranked first on two long-conversation memory benchmarks: achieving 93.8% recall@10 on LoCoMo and 65.9% on BEAM 100K against seven other embedding models. It also delivered the highest MRR and recall@5, highlighting its strength for retrieving relevant memories across long, complex conversation histories.

Perplexity saw strong results from NVIDIA Nemotron 3.5 Embed on the Q2D-Web benchmark. In their latest evaluation, NVIDIA Nemotron 3.5 Rerank 8B achieved the highest average score on the private 72-task reranking benchmark, with strong results across all six domains. These results highlight the value of specialized retrieval models alongside general-purpose decision models like Perplexity Decider family .

Turbopuffer evaluated NVIDIA Nemotron 3.5 Embed 8B on the enterprise PDF search benchmark and found that it provided ~5% more relevant results at Recall@20 than the next best model. These models are SOTA for common business tasks, like searching through PDFs with complex images, and are now available in Turbopuffer native embedding.

Getting Started

Developers can access the open weights, data, training recipes, and more on Hugging Face, deploy the NVIDIA NIM for production inference, or deploy the models through leading AI Cloud and Inference Partners including Baseten (Embed , Rerank , Parse), DeepInfra (Embed 1B , Embed 8B , Rerank 8B), and FriendlyAI (Embed 1B , Embed 8B , Rerank 8B).

Nemotron-3.5-Embed-8B-BF16: Hugging Face , NIM

Nemotron-3.5-Embed-8B-BF16: Hugging Face , NIM

Nemotron-3.5-Embed-1B-BF16: Hugging Face , NIM

Nemotron-3.5-Embed-1B-NVFP4: Hugging Face , NIM

Nemotron-3.5-Rerank-8B-BF16: Hugging Face , NIM

For domain adaptation with NVIDIA AutoModel, we provide fine-tuning support for multimodal embedding and reranking, enabling teams to adapt the models to their own document collections and relevance requirements. Get started with the Nemotron embedding and reranking fine-tuning recipes.

Choose a model based on your workflow: Parse for extracting structured information from documents, Embed for searching text and images, and Rerank for improving the ordering of retrieved results. Combine them to build retrieval pipelines that deliver stronger evidence to AI agents.

Acknowledgments and main contributors

Nemotron 3.5 Retriever release is the result of amazing collaboration between many teams and people across NVIDIA, including main contributors

Ronay Ak, Yauhen Babakhin, Behzad Boroujerdian, Jiarui Cai, Ryan Chesler, Zoe Faul, Jennifer Hajduk, Steve Han, Oliver Holworthy, Isabel Hulseman, Bo Liu, Gabriel Moreira, Even Oldridge, Radek Osmulski, Denis Ovchinnikov, Julio Perez, Khang Pham, Alex Richards, James Rong, Jihyeon Ryu, Sohail Sahi, Regis Schiavi, Sean Sodha, Marek Wawrzos, Jakub Zakrzewski, Lucy Zhang, Yuhe Zhang.

Models mentioned in this article 9

Datasets mentioned in this article 1

Spaces mentioned in this article 1

Collections mentioned in this article 1

More from this author

Announcing SimReady Central: An Open Library of Simulation-Ready 3D Assets for Robotics and AI Factories

1 October 21, 2026

NVIDIA Cosmos3-H-Dreams: Bringing Clinical Surgery to Life Through Generative Simulation

9 October 20, 2026

Community

Edit Preview

Upload images, audio, and videos by dragging in the text input, pasting, or clicking here . Tap or paste here to upload images

Comment · Sign up or log in to comment

Upvote 9

Models mentioned in this article 9

Datasets mentioned in this article 1

Spaces mentioned in this article 1

Collections mentioned in this article 1