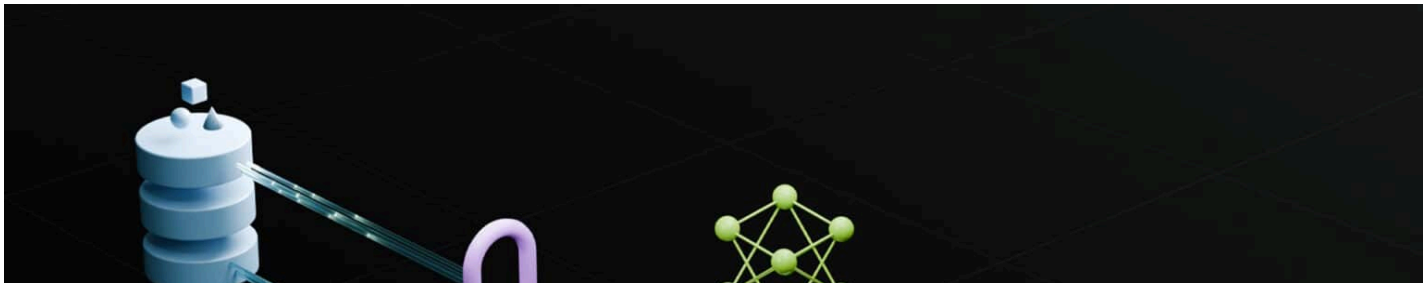
What Is Retrieval-Augmented Generation, aka RAG?

Retrieval-augmented generation is a technique for enhancing the accuracy and reliability of generative AI models with information from specific and relevant data sources.

January 31, 2025 by Rick Merritt



Share

Reading Time: 6 mins

Editor's note: This article, originally published on Nov. 15, 2023, has been updated.

To understand the latest advancements in generative AI, imagine a courtroom.

Judges hear and decide cases based on their general understanding of the law. Sometimes a case — like a malpractice suit or a labor dispute — requires special expertise, so judges send court clerks to a law library, looking for precedents and specific cases they can cite.

Like a good judge, large language models (LLMs) can respond to a wide variety of human queries. But to deliver authoritative answers — grounded in specific court proceedings or similar ones — the model needs to be provided that information.

The court clerk of AI is a process called retrieval-augmented generation, or RAG for short.

How It Got Named 'RAG'

Patrick Lewis, lead author of the 2020 paper that coined the term, apologized for the unflattering acronym that now describes a growing family of methods across hundreds of papers and dozens of commercial services he believes represent the future of generative AI.

Patrick Lewis

"We definitely would have put more thought into the name had we known our work would become so widespread," Lewis said in an interview from Singapore, where he was sharing his ideas with a regional conference of database developers.

"We always planned to have a nicer sounding name, but when it came time to write the paper, no one had a better idea," said Lewis, who now leads a RAG team at AI startup Cohere.

So, What Is Retrieval-Augmented Generation (RAG)?

Retrieval-augmented generation is a technique for enhancing the accuracy and reliability of generative AI models with information fetched from specific and relevant data sources.

In other words, it fills a gap in how LLMs work. Under the hood, LLMs are neural networks, typically measured by how many parameters they contain. An LLM's parameters essentially represent the general patterns of how humans use words to form sentences.

That deep understanding, sometimes called parameterized knowledge, makes LLMs useful in responding to general prompts. However, it doesn't serve users who want a deeper dive into a specific type of information.

Combining Internal, External Resources

Lewis and colleagues developed retrieval-augmented generation to link generative AI services to external resources, especially ones rich in the latest technical details.

The paper, with coauthors from the former Facebook AI Research (now Meta AI), University College London and New York University, called RAG "a general-purpose fine-tuning recipe" because it can be used by nearly any LLM to connect with practically any external resource.

Building User Trust

Retrieval-augmented generation gives models sources they can cite, like footnotes in a research paper, so users can check any claims. That builds trust.

What's more, the technique can help models clear up ambiguity in a user query. It also reduces the possibility that a model will give a very plausible but incorrect answer, a phenomenon called hallucination.

Another great advantage of RAG is it's relatively easy. A blog by Lewis and three of the paper's coauthors said developers can implement the process with as few as five lines of code.

That makes the method faster and less expensive than retraining a model with additional datasets. And it lets users hot-swap new sources on the fly.

How People Are Using RAG

With retrieval-augmented generation, users can essentially have conversations with data repositories, opening up new kinds of experiences. This means the applications for RAG could be multiple times the number of available datasets.

For example, a generative AI model supplemented with a medical index could be a great assistant for a doctor or nurse. Financial analysts would benefit from an assistant linked to market data.

In fact, almost any business can turn its technical or policy manuals, videos or logs into resources called knowledge bases that can enhance LLMs. These sources can enable use cases such as customer or field support, employee training and developer productivity.

The broad potential is why companies including AWS, IBM, Glean, Google, Microsoft, NVIDIA, Oracle and Pinecone are adopting RAG.

Getting Started With Retrieval-Augmented Generation

The NVIDIA AI Blueprint for RAG helps developers build pipelines to connect their AI applications to enterprise data using industry-leading technology. This reference architecture provides developers with a foundation for building scalable and customizable retrieval pipelines that deliver high accuracy and throughput.

The blueprint can be used as is, or combined with other NVIDIA Blueprints for advanced use cases including digital humans and AI assistants. For example, the blueprint for AI assistants empowers organizations to build AI agents that can quickly scale their customer service operations with generative AI and RAG.

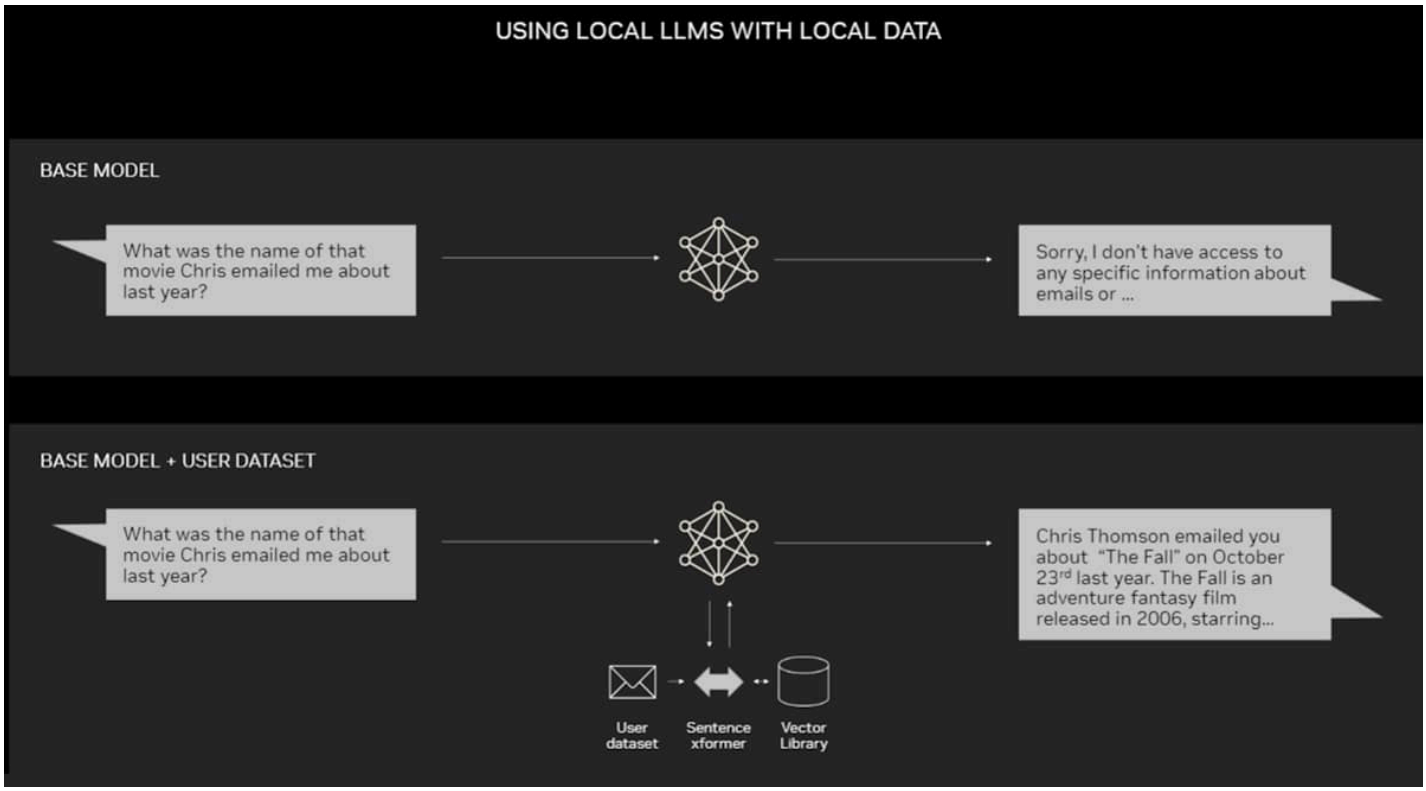
In addition, developers and IT teams can try the free, hands-on NVIDIA LaunchPad lab for building AI chatbots with RAG, enabling fast and accurate responses from enterprise data.

All of these resources use NVIDIA NeMo Retriever, which provides leading, large-scale retrieval accuracy and NVIDIA NIM microservices for simplifying secure, high-performance AI deployment across clouds, data centers and workstations. These are offered as part of the NVIDIA AI Enterprise software platform for accelerating AI development and deployment.

Getting the best performance for RAG workflows requires massive amounts of memory and compute to move and process data. The NVIDIA GH200 Grace Hopper Superchip, with its 288GB of fast HBM3e memory and 8 petaflops of compute, is ideal — it can deliver a 150x speedup over using a CPU.

Once companies get familiar with RAG, they can combine a variety of off-the-shelf or custom LLMs with internal or external knowledge bases to create a wide range of assistants that help their employees and customers.

RAG doesn’t require a data center. LLMs are debuting on Windows PCs, thanks to NVIDIA software that enables all sorts of applications users can access even on their laptops.



An example application for RAG on a PC.

PCs equipped with NVIDIA RTX GPUs can now run some AI models locally. By using RAG on a PC, users can link to a private knowledge source – whether that be emails, notes or articles – to improve responses. The user can then feel confident that their data source, prompts and response all remain private and secure.

A recent blog provides an example of RAG accelerated by TensorRT-LLM for Windows to get better results fast.

The History of RAG

The roots of the technique go back at least to the early 1970s. That's when researchers in information retrieval prototyped what they called question-answering systems, apps that use natural language processing (NLP) to access text, initially in narrow topics such as baseball.

The concepts behind this kind of text mining have remained fairly constant over the years. But the machine learning engines driving them have grown significantly, increasing their usefulness and popularity.

In the mid-1990s, the Ask Jeeves service, now Ask.com, popularized question answering with its mascot of a well-dressed valet. IBM's Watson became a TV celebrity in 2011 when it handily beat two human champions on the *Jeopardy!* game show.

Today, LLMs are taking question-answering systems to a whole new level.

Insights From a London Lab

The seminal 2020 paper arrived as Lewis was pursuing a doctorate in NLP at University College London and working for Meta at a new London AI lab. The team was searching for ways to pack more knowledge into an LLM's parameters and using a benchmark it developed to measure its progress.

Building on earlier methods and inspired by a paper from Google researchers, the group "had this compelling vision of a trained system that had a retrieval index in the middle of it, so it could learn and generate any text output you wanted," Lewis recalled.

The IBM Watson question-answering system became a celebrity when it won big on the TV game show Jeopardy!

When Lewis plugged into the work in progress a promising retrieval system from another Meta team, the first results were unexpectedly impressive.

“I showed my supervisor and he said, ‘Whoa, take the win. This sort of thing doesn’t happen very often,’ because these workflows can be hard to set up correctly the first time,” he said.

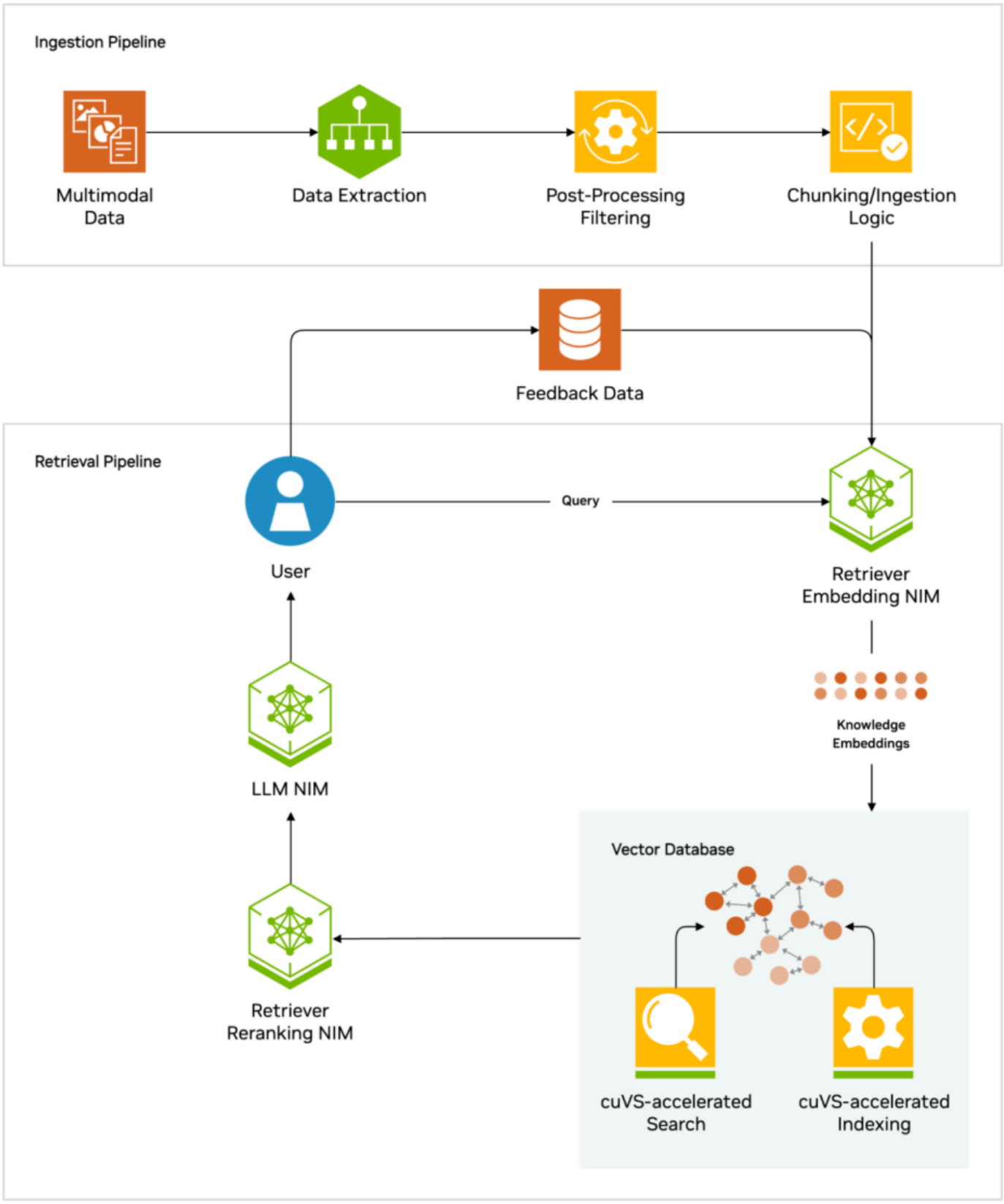
Lewis also credits major contributions from team members Ethan Perez and Douwe Kiela, then of New York University and Facebook AI Research, respectively.

When complete, the work, which ran on a cluster of NVIDIA GPUs, showed how to make generative AI models more authoritative and trustworthy. It’s since been cited by hundreds of papers that amplified and extended the concepts in what continues to be an active area of research.

How Retrieval-Augmented Generation Works

At a high level, here’s how retrieval-augmented generation works.

When users ask an LLM a question, the AI model sends the query to another model that converts it into a numeric format so machines can read it. The numeric version of the query is sometimes called an embedding or a vector.



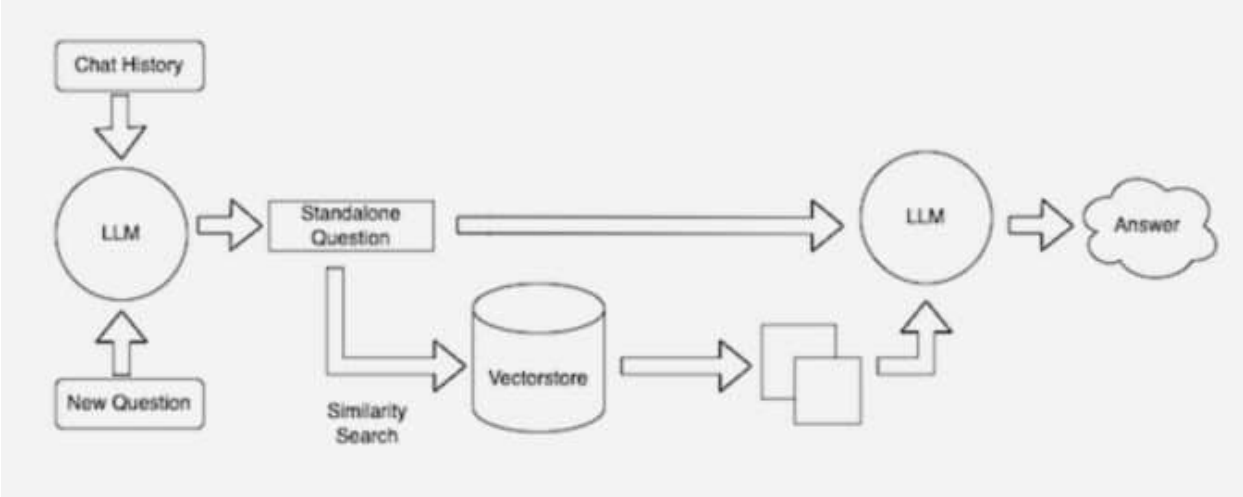
In retrieval-augmented generation, LLMs are enhanced with embedding and reranking models, storing knowledge in a vector database for precise query retrieval.

The embedding model then compares these numeric values to vectors in a machine-readable index of an available knowledge base. When it finds a match or multiple matches, it retrieves the related data, converts it to human-readable words and passes it back to the LLM.

Finally, the LLM combines the retrieved words and its own response to the query into a final answer it presents to the user, potentially citing sources the embedding model found.

Keeping Sources Current

In the background, the embedding model continuously creates and updates machine-readable indices, sometimes called vector databases, for new and updated knowledge bases as they become available.



Retrieval-augmented generation combines LLMs with embedding models and vector databases.

Many developers find LangChain, an open-source library, can be particularly useful in chaining together LLMs, embedding models and knowledge bases. NVIDIA uses LangChain in its reference architecture for retrieval-augmented generation.

The LangChain community provides its own description of a RAG process.

The future of generative AI lies in agentic AI — where LLMs and knowledge bases are dynamically orchestrated to create autonomous assistants. These AI-driven agents can enhance decision-making, adapt to complex tasks and deliver authoritative, verifiable results for users.

Categories: Deep Learning | Explainer | Generative AI

Tags: Artificial Intelligence | Events | Inference | Machine Learning | New GPU Uses | NVIDIA Blueprints | NVIDIA NeMo | TensorRT | Trustworthy AI

All NVIDIA News

Bring Receipts: New NVIDIA AI Blueprint Detects Fraudulent Credit Card Transactions With Precision

Researchers and Students in Türkiye Build AI, Robotics Tools to Boost Disaster Readiness

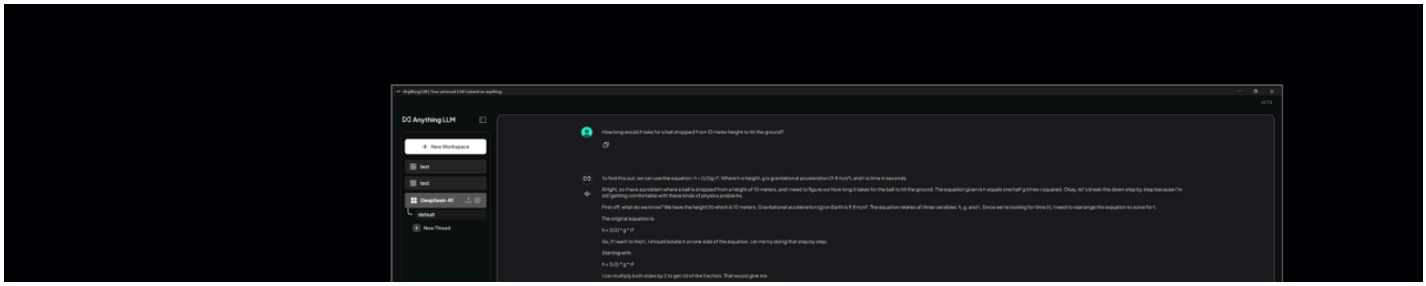
The More You Buy, the More You Make

The Supercomputer Designed to Accelerate Nobel-Worthy Science

RTX on Deck: The GeForce NOW Native App for Steam Deck Is Here

Accelerate DeepSeek Reasoning Models With NVIDIA GeForce RTX 50 Series AI PCs

January 31, 2025 by Annamalai Chockalingam



Share

Reading Time: 2 mins

The recently released DeepSeek-R1 model family has brought a new wave of excitement to the AI community, allowing enthusiasts and developers to run state-of-the-art reasoning models with problem-solving, math and code capabilities, all from the privacy of local PCs.

With up to 3,352 trillion operations per second of AI horsepower, NVIDIA GeForce RTX 50 Series GPUs can run the DeepSeek family of distilled models faster than anything on the PC market.

A New Class of Models That Reason

Reasoning models are a new class of large language models (LLMs) that spend more time on “thinking” and “reflecting” to work through complex problems, while describing the steps required to solve a task.

The fundamental principle is that any problem can be solved with deep thought, reasoning and time, just like how humans tackle problems. By spending more time — and thus compute — on a problem, the LLM can yield better results. This phenomenon is known as test-time scaling, where a model dynamically allocates compute resources during inference to reason through problems.

Reasoning models can enhance user experiences on PCs by deeply understanding a user’s needs, taking actions on their behalf and allowing them to provide feedback on the model’s thought process — unlocking agentic workflows for solving complex, multi-step tasks such as analyzing market research, performing complicated math problems, debugging code and more.

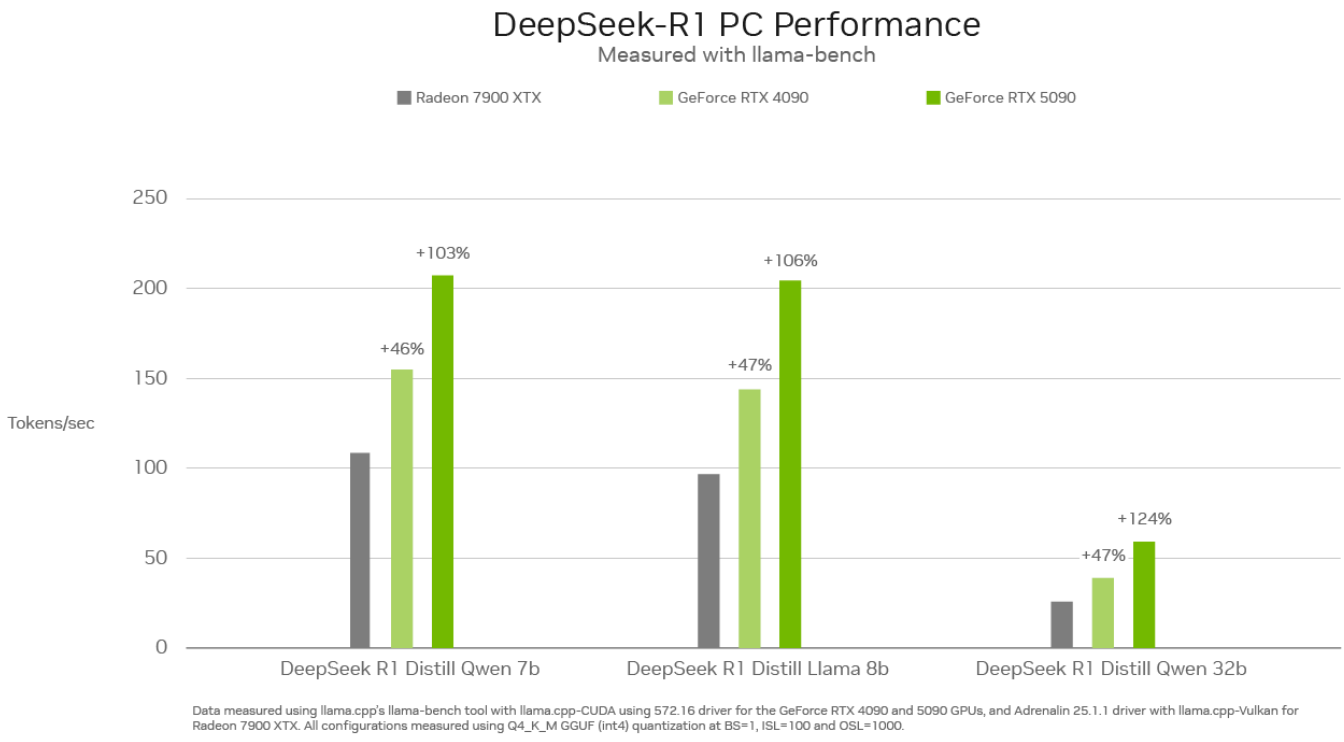
The DeepSeek Difference

The DeepSeek-R1 family of distilled models is based on a large 671-billion-parameter mixture-of-experts (MoE) model. MoE models consist of multiple smaller expert models for solving complex problems. DeepSeek models further divide the work and assign subtasks to smaller sets of experts.

DeepSeek employed a technique called distillation to build a family of six smaller student models — ranging from 1.5-70 billion parameters — from the large DeepSeek 671-billion-parameter model. The reasoning capabilities of the larger DeepSeek-R1 671-billion-parameter model were taught to the smaller Llama and Qwen student models, resulting in powerful, smaller reasoning models that run locally on RTX AI PCs with fast performance.

Peak Performance on RTX

Inference speed is critical for this new class of reasoning models. GeForce RTX 50 Series GPUs, built with dedicated fifth-generation Tensor Cores, are based on the same NVIDIA Blackwell GPU architecture that fuels world-leading AI innovation in the data center. RTX fully accelerates DeepSeek, offering maximum inference performance on PCs.



Throughput performance of the Deepseek-R1 distilled family of models across GPUs on the PC.

Experience DeepSeek on RTX in Popular Tools

NVIDIA's RTX AI platform offers the broadest selection of AI tools, software development kits and models, opening access to the capabilities of DeepSeek-R1 on over 100 million NVIDIA RTX AI PCs worldwide, including those powered by GeForce RTX 50 Series GPUs.

High-performance RTX GPUs make AI capabilities always available — even without an internet connection — and offer low latency and increased privacy because users don't have to upload sensitive materials or expose their queries to an online service.

Experience the power of DeepSeek-R1 and RTX AI PCs through a vast ecosystem of software, including Llama.cpp, Ollama, LM Studio, AnythingLLM, Jan.AI, GPT4All and OpenWebUI, for inference. Plus, use Unsloth to fine-tune the models with custom data.

Categories: Generative AI

Tags: Artificial Intelligence | NVIDIA RTX

All NVIDIA News

How 1X Technologies' Robots Are Learning to Lend a Helping Hand

NVIDIA Blackwell Delivers Breakthrough Performance in Latest MLPerf Training Results

NVIDIA RTX Blackwell GPUs Accelerate Professional-Grade Video Editing

Bring Receipts: New NVIDIA AI Blueprint Detects Fraudulent Credit Card Transactions With Precision

Researchers and Students in Türkiye Build AI, Robotics Tools to Boost Disaster Readiness

Mission Possible: Andres Diaz-Pinto Transforms Medical Imaging With MONAI

January 31, 2025 by Samantha Unnikrishnan

Reading Time: 2 mins

Andres Diaz-Pinto was among those at the forefront of transforming healthcare with AI well before he joined NVIDIA.

During his postdoctoral research in active learning for medical imaging at King's College London (KCL), Diaz-Pinto played an important role in creating MONAI Label, an open-source tool that uses AI to accelerate medical image annotation. It's part of the MONAI open-source project for AI in medical imaging that was pioneered by KCL and NVIDIA.

After a year of collaboration with the NVIDIA healthcare team on MONAI Label, Diaz-Pinto joined the company, bringing his expertise in developing and deploying AI and deep learning models in the cloud for medical image analysis.

Since joining NVIDIA in 2022 as a senior deep learning engineer, Diaz-Pinto has channeled his passion for research into his work.

“I work with some amazing people at NVIDIA, and there’s so much to learn from a technical perspective,” he said. “At the same time, I can continue my research work with KCL, other universities and hospitals.”

A native of Cúcuta, Colombia, Diaz-Pinto understood from an early age the importance of education. He became the first in his family to graduate high school and attend university. Today, he holds a doctorate in computer vision and deep learning and has completed two postdoctoral positions.

“Having faced hardships in Cúcuta, I quickly understood how important higher education would be in helping me and my family improve our situation,” he said.

In his current role, Diaz-Pinto helped advance brain surgery. The machine learning algorithms and computer vision technologies he’s helping create enhance the accuracy, speed and efficiency of medical imaging. And that allows radiologists to analyze medical images so they can better detect, diagnose and treat diseases, helping improve patient outcomes and reduce healthcare costs.

While he loves research, prioritizing life outside of work is just as important for Diaz-Pinto, especially as he’s a new father.

"My son was born 18 months ago, and I was able to spend lots of quality time with him thanks to NVIDIA's 12-week paternity leave policy," he said.

Fatherhood has given Diaz-Pinto a fresh perspective on life, deepening his love for learning. He also emphasizes the importance of hard work and humility that being a parent involves, saying, "You won't always have the best ideas or all the answers."

Outside of work and family, Diaz-Pinto enjoys cycling and exploring the city. During the recent company free days, he completed a 90-kilometer ride from London to Brighton in just four hours.

Always optimistic about the future, Diaz-Pinto believes AI is rapidly becoming a transformative force in healthcare. He envisions a future where robots are trained to enhance surgeons' dexterity to perform safer and faster operations on patients.

"At NVIDIA, I learn something new every day," he said. "Right now, I'm using the NVIDIA Omniverse platform to simulate surgical tasks with photorealistic digital twins of human organs."

Follow @nvidialife on Instagram and learn more about NVIDIA life, culture and careers.

Categories: NVIDIA Life

All NVIDIA News

How 1X Technologies' Robots Are Learning to Lend a Helping Hand

NVIDIA Blackwell Delivers Breakthrough Performance in Latest MLPerf Training Results

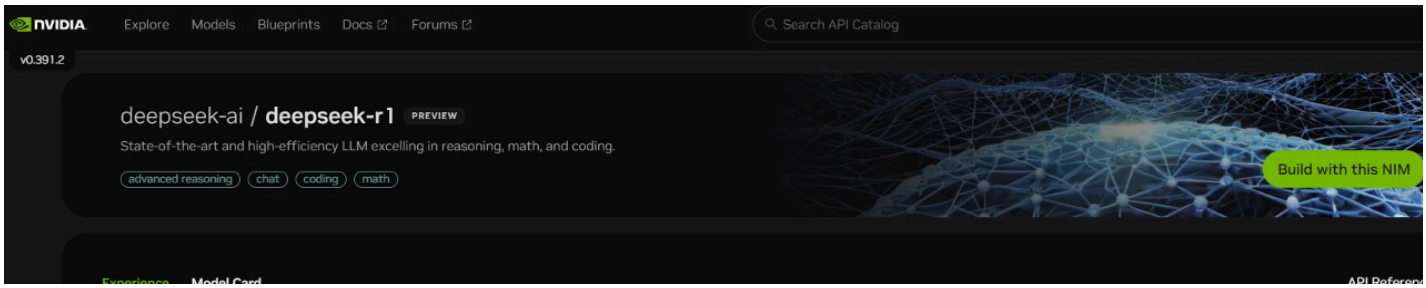
NVIDIA RTX Blackwell GPUs Accelerate Professional-Grade Video Editing

Bring Receipts: New NVIDIA AI Blueprint Detects Fraudulent Credit Card Transactions With Precision

Researchers and Students in Türkiye Build AI, Robotics Tools to Boost Disaster Readiness

DeepSeek-R1 Now Live With NVIDIA NIM

January 30, 2025 by Erik Pounds



Share

Reading Time: 3 mins

DeepSeek-R1 is an open model with state-of-the-art reasoning capabilities. Instead of offering direct responses, AI models like DeepSeek-R1 perform reasoning through the chain-of-thought method to generate the best answer.

Performing this sequence of inference passes — using reason to arrive at the best answer — is known as test-time scaling. DeepSeek-R1 is a perfect example of this scaling law, demonstrating why accelerated computing is critical for the demands of agentic AI inference.

As models are allowed to iteratively “think” through the problem, they create more output tokens and longer generation cycles, so model quality continues to scale. Significant test-time compute is critical to enable both real-time inference and higher-quality responses from reasoning models like DeepSeek-R1, requiring larger inference deployments.

R1 delivers leading accuracy for tasks demanding logical inference, reasoning, math, coding and language understanding while also delivering high inference efficiency.

To help developers securely experiment with these capabilities and build their own specialized agents, the 671-billion-parameter DeepSeek-R1 model is now available as an NVIDIA NIM microservice on build.nvidia.com. The DeepSeek-R1 NIM microservice can deliver up to 3,872 tokens per second on a single NVIDIA HGX H200 system.

Developers can test and experiment with the application programming interface (API), which is expected to be available soon as a downloadable NIM microservice, part of the NVIDIA AI Enterprise software platform.

The DeepSeek-R1 NIM microservice simplifies deployments with support for industry-standard APIs. Enterprises can maximize security and data privacy by running the NIM microservice on their preferred accelerated computing infrastructure. Using NVIDIA AI Foundry with NVIDIA NeMo software, enterprises will also be able to create customized DeepSeek-R1 NIM microservices for specialized AI agents.

DeepSeek-R1 — a Perfect Example of Test-Time Scaling

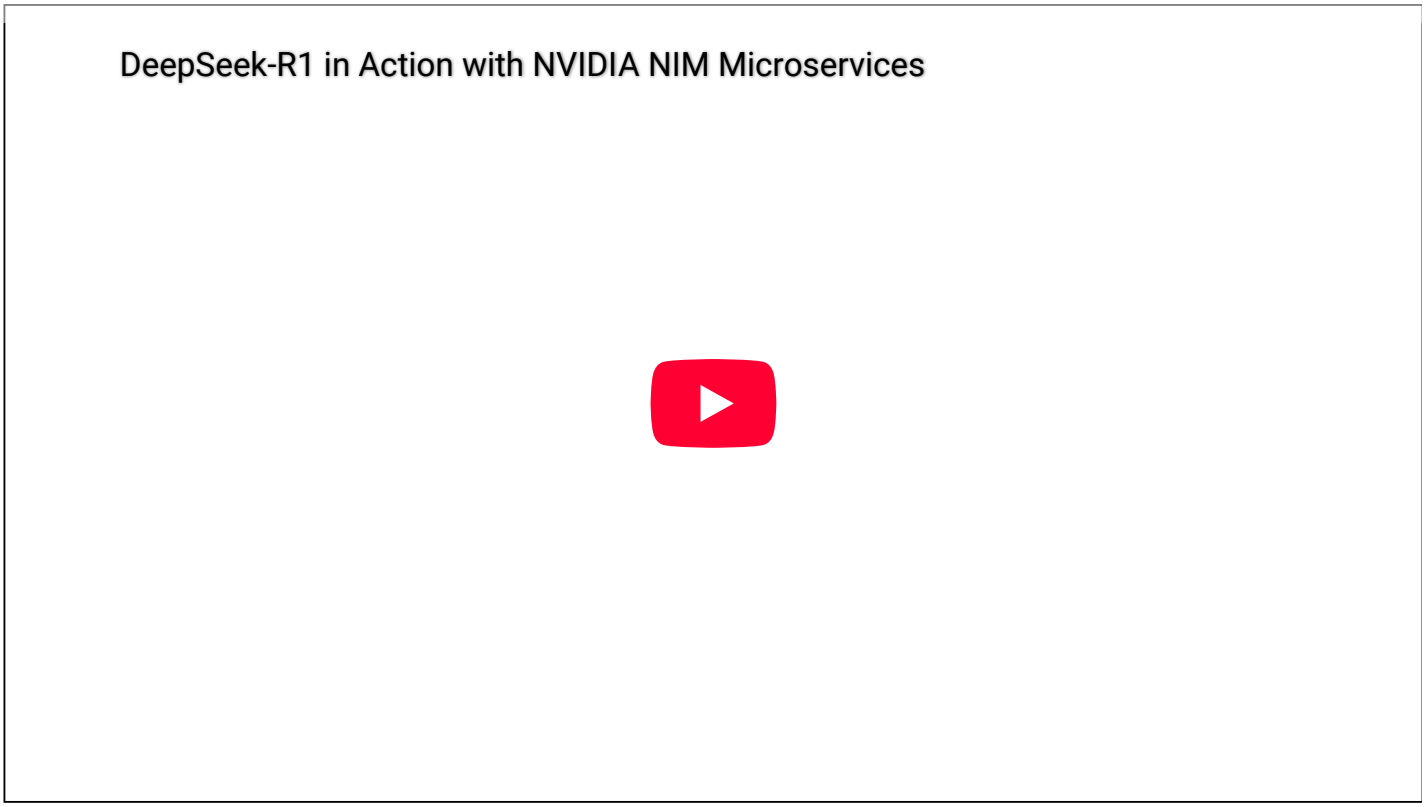
DeepSeek-R1 is a large mixture-of-experts (MoE) model. It incorporates an impressive 671 billion parameters — 10x more than many other popular open-source LLMs — supporting a large input context length of 128,000 tokens. The model also uses an extreme number of experts per layer. Each layer of R1 has 256 experts, with each token routed to eight separate experts in parallel for evaluation.

Delivering real-time answers for R1 requires many GPUs with high compute performance, connected with high-bandwidth and low-latency communication to route prompt tokens to all the experts for inference. Combined with the software optimizations available in the NVIDIA NIM microservice, a single server with eight H200 GPUs connected using NVLink and NVLink Switch can run the full, 671-billion-parameter DeepSeek-R1 model at up to 3,872 tokens per second. This throughput is made possible by using the NVIDIA Hopper architecture's FP8 Transformer Engine at every layer — and the 900 GB/s of NVLink bandwidth for MoE expert communication.

Getting every floating point operation per second (FLOPS) of performance out of a GPU is critical for real-time inference. The next-generation NVIDIA Blackwell architecture will give test-time scaling on reasoning models like DeepSeek-R1 a giant boost with fifth-generation Tensor Cores that can deliver up to 20 petaflops of peak FP4 compute performance and a 72-GPU NVLink domain specifically optimized for inference.

Get Started Now With the DeepSeek-R1 NIM Microservice

Developers can experience and download the DeepSeek-R1 NIM microservice, now in general availability on build.nvidia.com. Watch how it works:



With NVIDIA NIM, enterprises can deploy DeepSeek-R1 with ease and ensure they get the high efficiency needed for agentic AI systems.

See notice regarding software product information.

Categories: Generative AI

Tags: AI Agents | Artificial Intelligence | NVIDIA NeMo | NVIDIA NIM | NVLink

All NVIDIA News

How 1X Technologies' Robots Are Learning to Lend a Helping Hand

NVIDIA Blackwell Delivers Breakthrough Performance in Latest MLPerf Training Results

NVIDIA RTX Blackwell GPUs Accelerate Professional-Grade Video Editing

Bring Receipts: New NVIDIA AI Blueprint Detects Fraudulent Credit Card Transactions With Precision

Researchers and Students in Türkiye Build AI, Robotics Tools to Boost Disaster Readiness