

What Is Retrieval-Augmented Generation, aka RAG?


Retrieval-augmented generation (RAG) is a technique for enhancing the accuracy and reliability of generative AI models with facts fetched from external sources.

November 18, 2024 by [Rick Merritt](#)



 Share



 Feedback



Reading Time: 6 mins

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

To understand the latest advance 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, deliver authoritative answers that cite [sources](#), the model needs an assistant to do some research.

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 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 at 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 (RAG) is a technique for enhancing the accuracy and reliability of generative AI models by fetching facts from external 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 at light speed. However, it does not serve users who want a deeper dive into a current or more specific topic.

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 a model will make a wrong guess, a phenomenon sometimes 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 dataset 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 reposits opening up new kinds of experiences. This means the applications for RAG could be multiple time number of available datasets.

For example, a generative AI model supplemented with a medical index could be a great assistant 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 knowledge bases that can enhance LLMs. These sources can enable use cases such as customer field support, employee training and developer productivity.

The broad potential is why companies including [AWS](#), [IBM](#), [Glean](#), Google, Microsoft, NVIDIA, [Oracle](#), [Pinecone](#) are adopting RAG.

Getting Started With Retrieval-Augmented Generation

To help users get started, NVIDIA developed an [AI Blueprint](#) for building virtual assistants. Organizations can use this reference architecture to quickly scale their customer service operations with generative AI and RAG, or get started building a new customer-centric solution.

The blueprint uses some of the latest AI-building methodologies and [NVIDIA NeMo Retriever](#), a collection of easy-to-use [NVIDIA NIM](#) microservices for large-scale information retrieval. NIM enables secure deployment of secure, high-performance AI model inferencing across clouds, data centers and workstations.

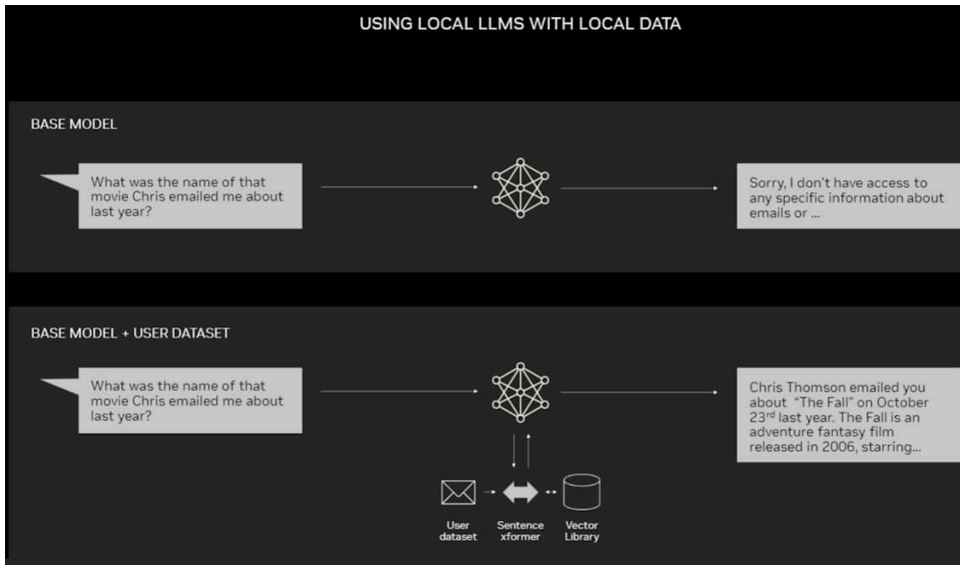
These components are all part of [NVIDIA AI Enterprise](#), a software platform that accelerates the development and deployment of production-ready AI with the security, support and stability businesses need.

There is also a free hands-on [NVIDIA LaunchPad lab](#) for developing AI chatbots using RAG so developers and IT teams can quickly and accurately generate responses based on enterprise data.

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, you 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.



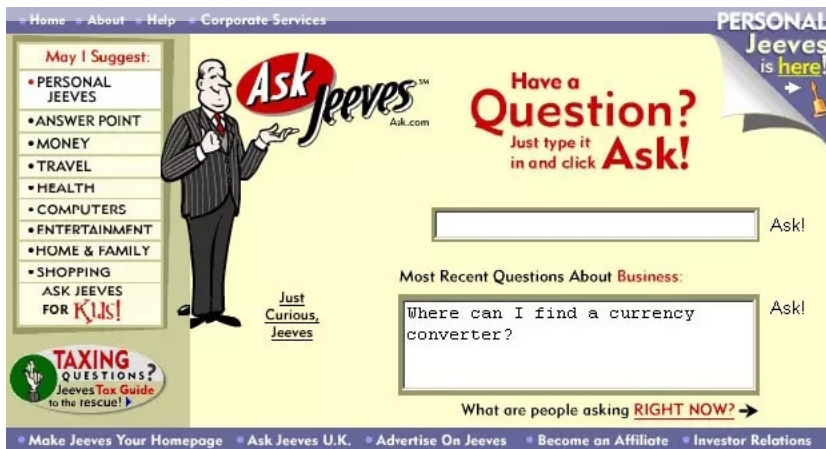
A recent [blog](#) provides an example of RAG accelerated by TensorRT-LLM for Windows to get better results faster.

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 role 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 Lc 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 to 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 te 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 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 York University and Facebook AI Research, respectively.

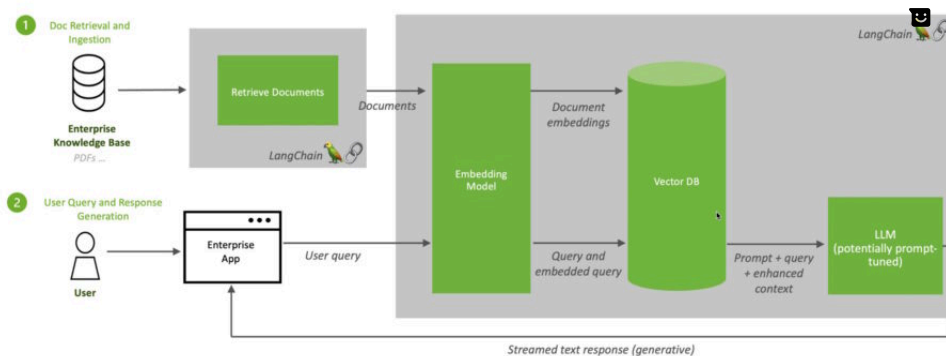
When complete, the work, which ran on a cluster of NVIDIA GPUs, showed how to make generativ models more authoritative and trustworthy. It's since been cited by hundreds of papers that amp and extended the concepts in what continues to be an active area of research.

How Retrieval-Augmented Generation Works

At a high level, here's how an NVIDIA technical brief describes the RAG process.

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

Retrieval Augmented Generation (RAG) Sequence Diagram



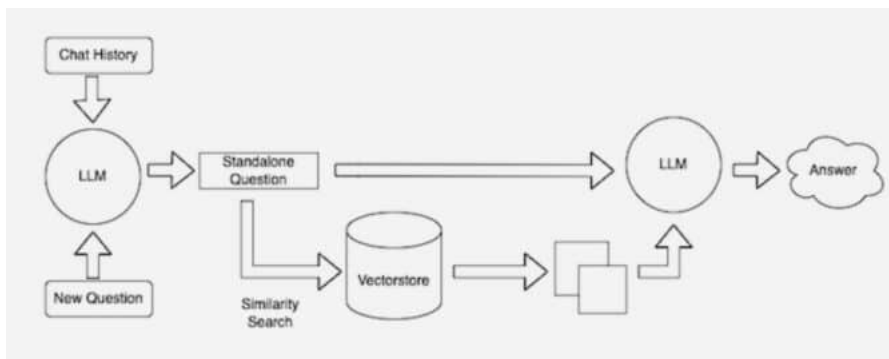
Retrieval-augmented generation combines LLMs with embedding models and vector databases.

The embedding model then compares these numeric values to vectors in a machine-readable ind an available knowledge base. When it finds a match or multiple matches, it retrieves the related d 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 answ 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 information, 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](#).

Looking forward, the future of generative AI lies in creatively chaining all sorts of LLMs and knowledge bases together to create new kinds of assistants that deliver authoritative results users can verify.

Explore [generative AI sessions and experiences at NVIDIA GTC](#), the global conference on AI and accelerated computing, running March 18-21 in San Jose, Calif., and online. 

Categories: [Deep Learning](#) | [Explainer](#) | [Generative AI](#)

Tags: [Artificial Intelligence](#) | [Events](#) | [Inference](#) | [Machine Learning](#) | [New GPU Uses](#) | [NVIDIA Blueprint](#) | [NVIDIA NeMo](#) | [TensorRT](#) | [Trustworthy AI](#)



What's Next in AI Starts Here

March 17–21, 2025

Get Early-Bird Pricing



Recommended for You

AI Opener: OpenAI's Sutskever in Conversation With Jensen Huang

Ubisoft's Yves Jacquier on How Generative AI Will Revolutionize Gaming

Blender Update 3.5 Fuels 3D Content Creation, Powered by NVIDIA GeForce RTX GPUs

Unreal Engine Enhances Virtual Production Workflows With NVIDIA Rivermax and BlueField

NVIDIA Honors Partners in Americas Helping Industries Harness AI to Transform Business

Stay up to date on the latest enterprise news.

Professional Email Address

Subscribe Now

Open for Development: NVIDIA Works With Cloud-Native Community to Advance AI and ML

At KubeCon+CloudNativeCon, NVIDIA engineers and experts showcase open-source software contributions through a keynote, technical sessions, training and demos.

November 14, 2024 by [Ankit Patel](#)



 Share



Reading Time: 3 mins

Cloud-native technologies have become crucial for developers to create and implement scalable applications in dynamic cloud environments.

This week at KubeCon + CloudNativeCon North America 2024, one of the most-attended conference focused on open-source technologies, Chris Lamb, vice president of computing software platform NVIDIA, delivered a keynote outlining the benefits of open source for developers and enterprises and NVIDIA offered nearly 20 interactive sessions with engineers and experts.

The Cloud Native Computing Foundation (CNCF), part of the Linux Foundation and host of Kubernetes at the forefront of championing a robust ecosystem to foster collaboration among industry leaders, developers and end users.

As a member of CNCF since 2018, NVIDIA is working across the developer community to contribute and sustain cloud-native open-source projects. Our open-source software and more than 750 NVIDIA-led open-source projects help democratize access to tools that accelerate AI development and innovation.

Empowering Cloud-Native Ecosystems

NVIDIA has benefited from the many open-source projects under CNCF and has made contributions to dozens of them over the past decade. These actions help developers as they build applications and microservice architectures aligned with managing AI and machine learning workloads.

Kubernetes, the cornerstone of cloud-native computing, is undergoing a transformation to meet the challenges of AI and machine learning workloads. As organizations increasingly adopt large language models and other AI technologies, robust infrastructure becomes paramount.

NVIDIA has been working closely with the Kubernetes community to address these challenges. This includes:

- Work on dynamic resource allocation (DRA) that allows for more flexible and nuanced resource management. This is crucial for AI workloads, which often require specialized hardware. NVIDIA engineers played a key role in designing and implementing this feature.
- Leading efforts in KubeVirt, an open-source project extending Kubernetes to manage virtual machines alongside containers. This provides a unified, cloud-native approach to managing hybrid infrastructure.
- Development of NVIDIA GPU Operator, which automates the lifecycle management of NVIDIA GPUs in Kubernetes clusters. This software simplifies the deployment and configuration of GPU drivers, runtime and monitoring tools, allowing organizations to focus on building AI applications rather than managing infrastructure.

The company's open-source efforts extend beyond Kubernetes to other CNCF projects:

- NVIDIA is a key contributor to KubeFlow, a comprehensive toolkit that makes it easier for data scientists and engineers to build and manage ML systems on Kubernetes. KubeFlow reduces the complexity of infrastructure management and allows users to focus on developing and improving models.
- NVIDIA has contributed to the development of CNAO, which manages the lifecycle of host networks in Kubernetes clusters.
- NVIDIA has also added to Node Health Check, which provides virtual machine high availability.

And NVIDIA has assisted with projects that address the observability, performance and other critical areas of cloud-native computing, such as:

- Prometheus: Enhancing monitoring and alerting capabilities
- Envoy: Improving distributed proxy performance

- OpenTelemetry: Advancing observability in complex, distributed systems
- Argo: Facilitating Kubernetes-native workflows and application management

Community Engagement

NVIDIA engages the cloud-native ecosystem by participating in CNCF events and activities, includ

- Collaboration with cloud service providers to help them onboard new workloads.
- Participation in CNCF's special interest groups and working groups on AI discussions.
- Participation in industry events such as KubeCon + CloudNativeCon, where it shares insights o acceleration for AI workloads.
- Work with CNCF-adjacent projects in the Linux Foundation as well as many partners.

This translates into extended benefits for developers, such as improved efficiency in managing AI ML workloads; enhanced scalability and performance of cloud-native applications; better resource utilization, which can lead to cost savings; and simplified deployment and management of complex infrastructures.

As AI and machine learning continue to transform industries, NVIDIA is helping advance cloud-native technologies to support compute-intensive workloads. This includes facilitating the migration of applications and supporting the development of new ones.

These contributions to the open-source community help developers harness the full potential of technologies and strengthen Kubernetes and other CNCF projects as the tools of choice for AI compute workloads.

Check out [NVIDIA's keynote at KubeCon + CloudNativeCon North America 2024](#) delivered by Christy La where he discusses the importance of CNCF projects in building and delivering AI in the cloud and NVIDIA contributions to the community to push the AI revolution forward.

Categories: [Cloud](#) | [Software](#)

Tags: [Events](#) | [Open Source](#) | [Social Impact](#)

Load Comments



What's Next in AI Starts Here

March 17–21, 2025

Get Early-Bird Pricing



Recommended for You

AI Opener: OpenAI's Sutskever in Conversation With Jensen Huang

Ubisoft's Yves Jacquier on How Generative AI Will Revolutionize Gaming

Blender Update 3.5 Fuels 3D Content Creation, Powered by NVIDIA GeForce RTX GPUs

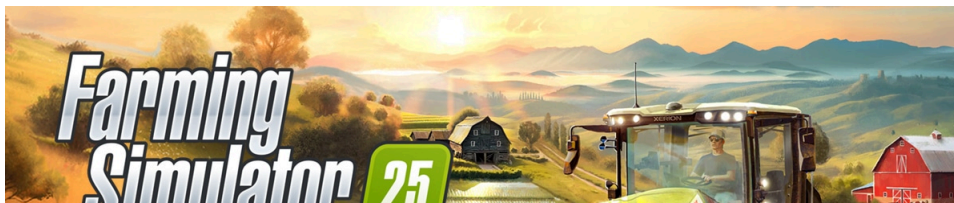
Unreal Engine Enhances Virtual Production Workflows With NVIDIA Rivermax and BlueField

NVIDIA Honors Partners in Americas Helping Industries Harness AI to Transform Business

From Seed to Stream: 'Farming Simulator Sprouts on GeForce NOW

A new PC-exclusive GeForce NOW reward for members in 'Throne and Liberty' means it's time to unleash some mischief; plus, nine games are arriving to the cloud.

November 14, 2024 by [GeForce NOW Community](#)



 Share



Reading Time: 4 mins

Grab a pitchfork and fire up the tractor — the fields of GeForce NOW are about to get a whole lot greener with *Farming Simulator 25*.

Whether looking for a time-traveling adventure, cozy games or epic action, [GeForce NOW](#) has something for everyone with over 2,000 games in its cloud library. Nine titles arrive this week, including the new 4X historical grand strategy game *Ara: History Untold* from Oxide Games and Xbox Game Studios.


And in this season of giving, GeForce NOW will offer members new rewards and more this month. This week, GeForce NOW is spreading cheer with a new reward for members that's sure to delight *Throne and Liberty* fans. Get ready to add a dash of mischief and a sprinkle of wealth to the epic adventure in the sprawling world of this massively multiplayer online role-playing game.

Plus, the [NVIDIA app](#) is officially released for download this week. GeForce users can use it to access GeForce NOW to play their games with RTX performance when they're away from their gaming rig or don't want to wait around for their games to update and patch.

A Cloud Gaming Bounty

Get ready to plow the fields and tend to crops anywhere with GeForce NOW.



Farming Simulator 25 from Giants Software launched in the cloud for members to stream, bringing a host of new features and improvements — including the introduction of rice as a crop type, complete with specialized machinery and techniques for planting, flooding fields and harvesting. 

This expansion into rice farming is accompanied by a new Asian-themed map that offers players a landscape filled with picturesque rice paddies to cultivate. The game will also include two other diverse environments: a spacious North American setting and a scenic Central European location, allowing farmers to build their agricultural empires in varied terrains. Don't forget about the addition of new animals like buffaloes and goats, as well as the introduction of animal offspring for a new layer of depth to farm management.

Be the cream of the crop streaming with a Performance or Ultimate membership. Performance members get up to 1440p 60 frames per second and Ultimate streams at up to 4K and 120 fps for the most incredible levels of realism and variety. Whether tackling agriculture, forestry and animal husbandry single-handedly or together with friends in cooperative multiplayer mode, experience farming life like never before with GeForce NOW.

Mischief Managed

Whether new to the game or a seasoned adventurer, GeForce NOW members can claim a special exclusive reward to use in Amazon Games' hit title *Throne and Liberty*. The reward includes 200 Orbs and Coins and a PC-exclusive mischievous youngster named Gneiss Amitoi that will enhance the *Throne and Liberty* journey as members forge alliances, wage epic battles and uncover hidden treasures.



Ornate Coins allow players to acquire morphs for animal shapeshifting, autonomous pets named Amitois, exclusive cosmetic items, experience boosters and inventory expansions. Gneiss Youngst Amitoi is a toddler-aged prankster that randomly targets players and non-playable characters with tricks. While some of its mischief can be mean-spirited, it just wants attention, and will pout and back to its adventurer's side if ignored, adding an entertaining dynamic to the journey through the world of *Throne and Liberty*.

Members who've opted in to GeForce NOW's Rewards program can check their email for instructions on how to redeem the reward. Ultimate and Performance members can start redeeming the reward now, while free members will be able to claim it starting tomorrow, Nov. 15. It's available through Tuesday, Dec. 10, first come, first served.

Rewriting History

Explore, build, lead and conquer a nation in *Ara: History Untold*, where every choice will shape the world and define a player's legacy. It's now available for GeForce NOW members to stream.

Ara: History Untold offers a fresh take on 4X historical grand strategy games. Players will prove their worth by guiding their citizens through history to the pinnacles of human achievement. Explore new lands, develop arts and culture, and engage in diplomacy — or combat — with other nations, before ultimately claiming the mantle of the greatest nation of all time.

Members can craft their own unique story of triumph and achievement by streaming the game on any devices from the cloud. GeForce NOW Performance and Ultimate members can enjoy longer game sessions and faster access to servers than free users, perfect for crafting sprawling empires and engaging in complex diplomacy without worrying about local hardware limitations.

New Games Are Knocking

Wuthering Waves Version 1.4 Official Trailer | When the Night Knocks



GeForce NOW brings the new *Wuthering Waves* update “When the Night Knocks” for members this week. Version 1.4 brings a wealth of new content, including two new Resonators, Camellya and Lu along with powerful new weapons, including the five-star Red Spring and the four-star event weapon Somnoire Anchor. Dive into the Somnoire Adventure Event, *Somnium Labyrinth*, and enjoy a variety of log-in rewards, combat challenges and exploration activities. The update also includes Camellya’s companion story, a new Phantom Echo and introduces the exciting Weapon Projection feature 😊

Members can look for the following games available to stream in the cloud this week:

- *Farming Simulator 25* (New release on [Steam](#), Nov. 12)
- *Sea Power: Naval Combat in the Missile Age* (New release on [Steam](#), Nov. 12)
- *Industry Giant 4.0* (New release [Steam](#), Nov. 15)
- *Ara: History Untold* ([Steam](#) and [Xbox](#), available on PC Game Pass)
- *Call of Duty: Black Ops Cold War* ([Steam](#) and [Battle.net](#))
- *Call of Duty: Vanguard* ([Steam](#) and [Battle.net](#))
- *Minecraft* ([Steam](#))
- *Crash Bandicoot N. Sane Trilogy* ([Steam](#) and [Xbox](#), available on PC Game Pass)
- *Spyro Reignited Trilogy* ([Steam](#) and [Xbox](#), available on PC Game Pass)

What are you planning to play this weekend? Let us know on [X](#) or in the comments below.

**NVIDIA GeForce NOW**  

@NVIDIAGFN · [Follow](#)

the last thing your left hand touched is your video game weapon - what was it?

10:44 PM · Nov 13, 2024 

 72

 [Reply](#)

 [Copy link](#)

[Read 58 replies](#)

Categories: [Corporate Sustainability](#) | [Gaming](#)

Tags: [Cloud Gaming](#) | [GeForce NOW](#)

Load Comments





What's Next in AI Starts Here

March 17–21, 2025

Get Early-Bird Pricing



Recommended for You

AI Opener: OpenAI's Sutskever in Conversation With Jensen Huang

Ubisoft's Yves Jacquier on How Generative AI Will Revolutionize Gaming

Blender Update 3.5 Fuels 3D Content Creation, Powered by NVIDIA GeForce RTX GPUs

Unreal Engine Enhances Virtual Production Workflows With NVIDIA Rivermax and BlueField

NVIDIA Honors Partners in Americas Helping Industries Harness AI to Transform Business

Keeping an AI on Diabetes Risk: Gen AI Model Predicts Blood Sugar Levels Four Years Out

November 14, 2024 by [Isha Salian](#)



 Share

f

in



Reading Time: 3 mins

Diabetics — or others monitoring their sugar intake — may look at a cookie and wonder, “How will this affect my glucose levels?” A generative AI model can now predict the answer.

Researchers from the Weizmann Institute of Science, Tel Aviv-based startup Pheno.AI and NVIDIA announced the development of [GluFormer](#), an AI model that can predict an individual's future glucose levels and other health metrics based on past glucose monitoring data.

Data from continuous glucose monitoring could help more quickly diagnose patients with prediabetes or diabetes, according to [Harvard Health Publishing](#) and [NYU Langone Health](#). GluFormer's AI capabilities can further enhance the value of this data, helping clinicians and patients spot anomalies, predict clinical trial outcomes and forecast health outcomes up to four years in advance.

The researchers showed that, after adding dietary intake data into the model, GluFormer can also predict how a person's glucose levels will respond to specific foods and dietary changes, enabling precision nutrition.

Accurate predictions of glucose levels for those at high risk of developing diabetes could enable doctors and patients to adopt preventative care strategies sooner, improving patient outcomes as reducing the economic impact of diabetes, which could reach [\\$2.5 trillion globally by 2030](#).

AI tools like GluFormer have the potential to help the hundreds of millions of adults with diabetes condition currently affects around 10% of the world's adults — a figure that could potentially [double by 2050](#) to impact over 1.3 billion people. It's one of the [10 leading causes of death globally](#), with side effects including kidney damage, vision loss and heart problems.

GluFormer is a [transformer model](#), a kind of neural network architecture that tracks relationships sequential data. It's the same architecture as OpenAI's GPT models — in this case generating glucose levels instead of text.

"Medical data, and continuous glucose monitoring in particular, can be viewed as sequences of diagnostic tests that trace biological processes throughout life," said Gal Chechik, senior director research at NVIDIA. "We found that the transformer architecture, developed for long text sequences, can take a sequence of medical tests and predict the results of the next test. In doing so, it learns something about how the diagnostic measurements develop over time."

The model was trained on 14 days of glucose monitoring data from over 10,000 non-diabetic study participants, with data collected every 15 minutes through a wearable monitoring device. The data collected as part of the [Human Phenotype Project](#), an initiative by Pheno.AI, a startup that aims to improve human health through data collection and analysis.

"Two important factors converged at the same time to enable this research: the maturing of generative AI technology powered by NVIDIA and the collection of large-scale health data by the Weizmann Institute," said the paper's lead author, Guy Lutsker, an NVIDIA researcher and Ph.D. student at the Weizmann Institute of Science. "It put us in the unique position to extract interesting medical insights from the data."

The research team validated GluFormer across 15 other datasets and found it generalizes well to predict health outcomes for other groups, including those with prediabetes, type 1 and type 2 diabetes, gestational diabetes and obesity.

They used a cluster of [NVIDIA Tensor Core GPUs](#) to accelerate model training and inference.

Beyond glucose levels, GluFormer can predict medical values including visceral adipose tissue, a measure of the amount of body fat around organs like the liver and pancreas; systolic blood pressure, which is associated with diabetes risk; and apnea-hypopnea index, a measurement for sleep apnea, which is linked to type 2 diabetes.

Read the [GluFormer research paper on Arxiv](#).

Categories: [Generative AI](#) | [Research](#)

Tags: [Artificial Intelligence](#) | [Healthcare and Life Sciences](#) | [NVIDIA Research](#) | [Science](#) | [Social Impact](#)

[Load Comments](#)



What's Next in AI Starts Here

March 17–21, 2025

Get Early-Bird Pricing



Recommended for You

AI Opener: OpenAI's Sutskever in Conversation With Jensen Huang

Ubisoft's Yves Jacquier on How Generative AI Will Revolutionize Gaming

Blender Update 3.5 Fuels 3D Content Creation, Powered by NVIDIA GeForce RTX GPUs

Unreal Engine Enhances Virtual Production Workflows With NVIDIA Rivermax and BlueField

NVIDIA Honors Partners in Americas Helping Industries Harness AI to Transform Business

