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 15, 2023 by Rick Merritt

Share

Reading Time: 6 mins

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. But to 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 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 (RAG) is a technique for enhancing the accuracy and reliability of generative AI models with facts fetched 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 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

To help users get started, NVIDIA developed an AI workflow for retrieval-augmented generation. It includes a sample chatbot and the elements users need to create their own applications with this new method.

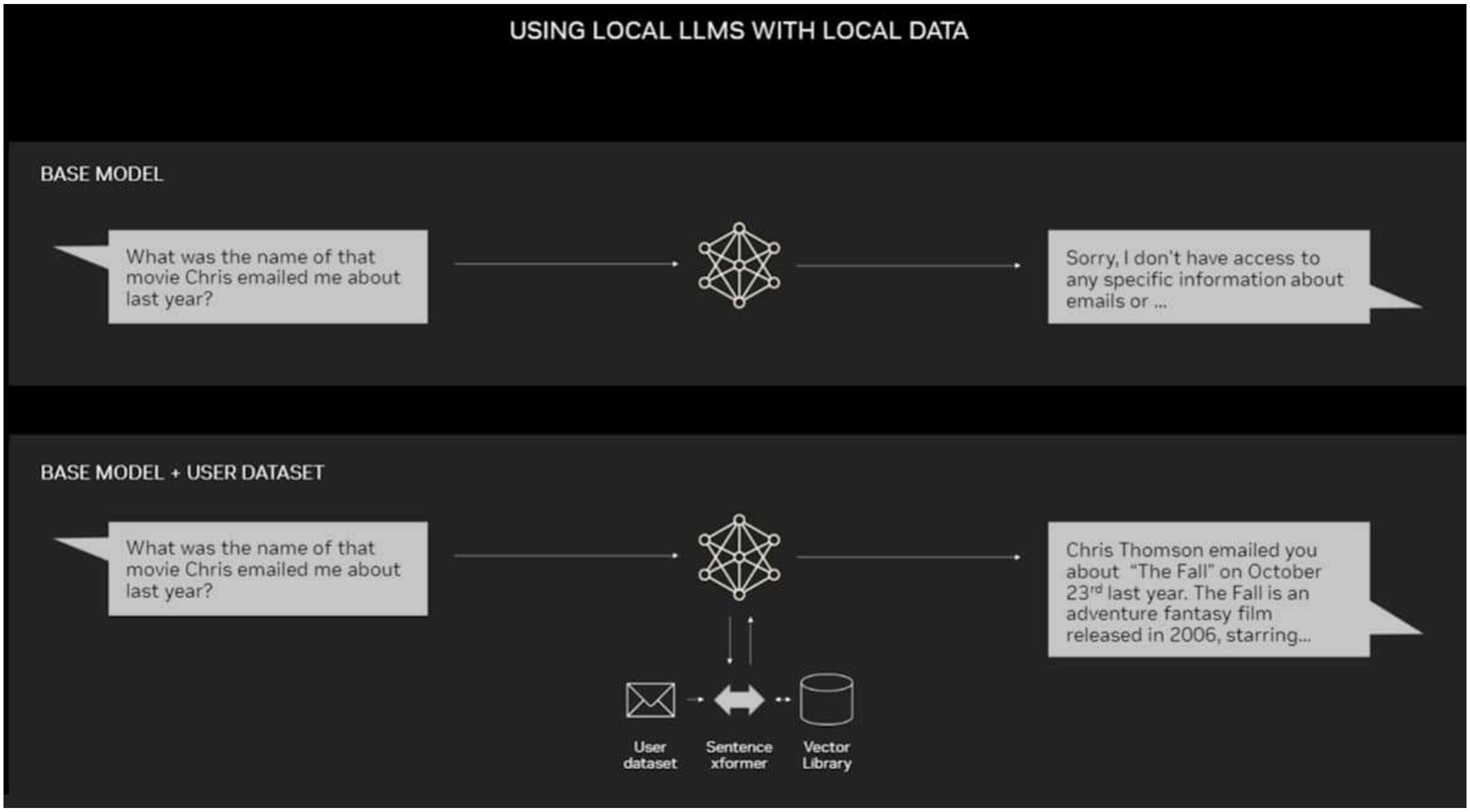
The workflow uses NVIDIA NeMo, a framework for developing and customizing generative AI models, as well as software like NVIDIA Triton Inference Server and NVIDIA TensorRT-LLM for running generative AI models in production.

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

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.


USING LOCAL LLMS WITH LOCAL DATA
BASE MODEL
What was the name of that movie Chris emailed me about last year?
Sorry, I don't have access to any specific information about emails or ...
BASE MODEL + USER DATASET
What was the name of that movie Chris emailed me about last year?
Chris Thomson emailed you about "The Fall" on October 23rd last year. The Fall is an adventure fantasy film released in 2006, starring...
User dataset
Sentence xformer
Vector Library

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.

Where can I find a currency converter?

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.

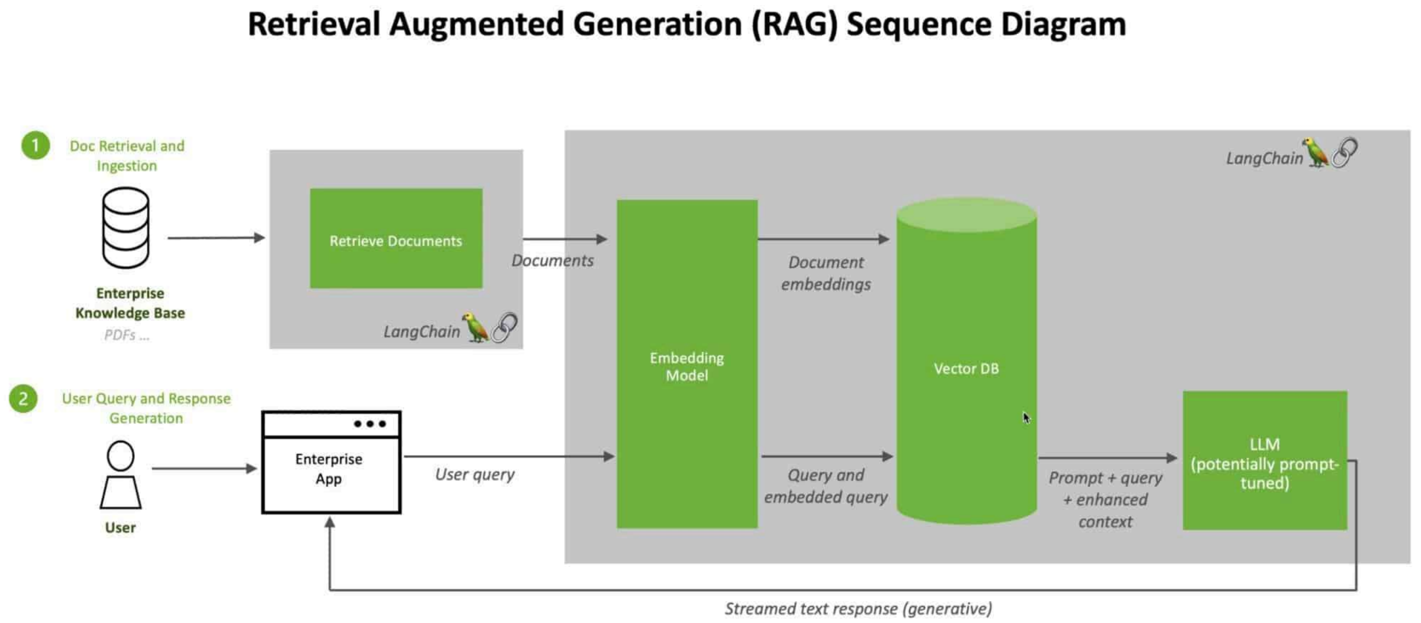
Lewis also credits major contributions from team members Ethan Perez and Douwe Kiela, then of New York University and Facebook AI Research, respectively.

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 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 it into a numeric format so machines can read it. The numeric version of the query is sometimes called an embedding or a vector.


Retrieval Augmented Generation (RAG) Sequence Diagram
1 Doc Retrieval and Ingestion
Enterprise Knowledge Base
PDFs ...
Retrieve Documents
LangChain
Documents
Embedding Model
Document embeddings
Vector DB
LangChain
2 User Query and Response Generation
User
Enterprise App
User query
Query and embedded query
Prompt + query + enhanced context
LLM (potentially prompt-tuned)
Streamed text response (generative)

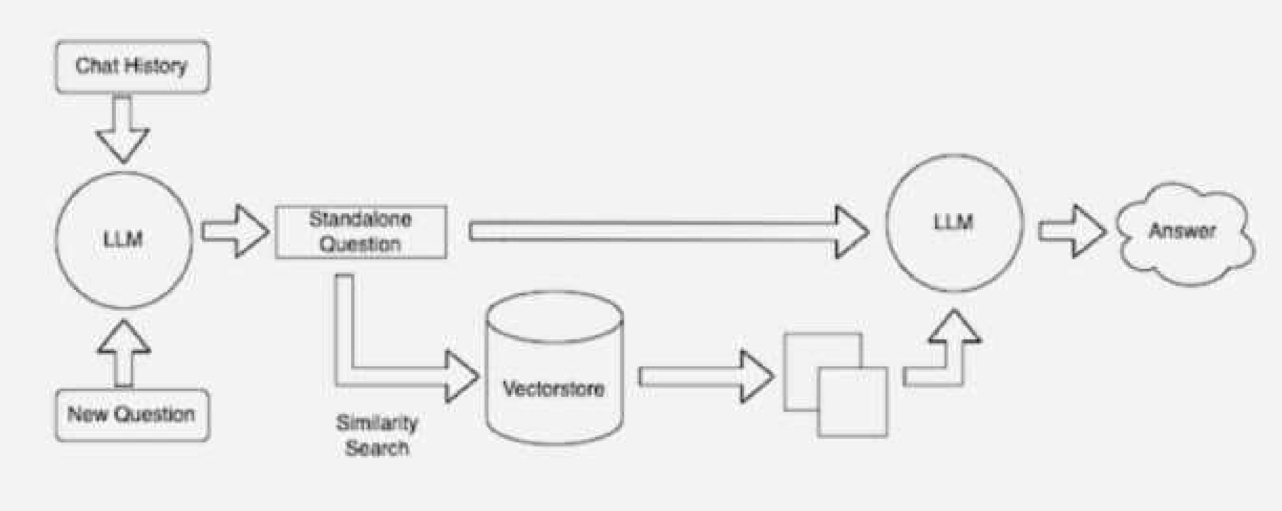
Retrieval-augmented generation combines LLMs with embedding models and vector databases.

The embedding model then compares these numeric values to vectors in a machine-readable 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.

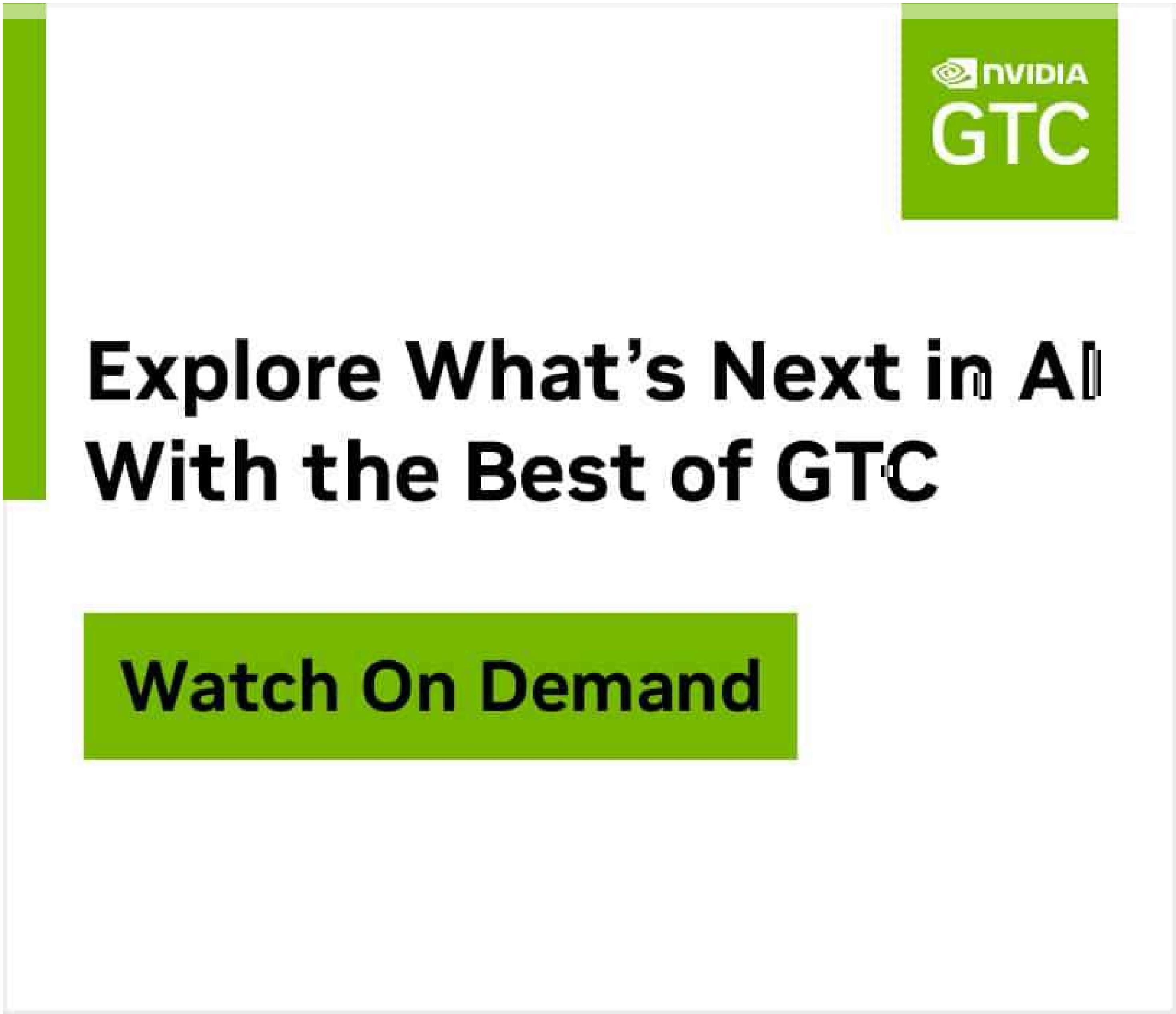
Looking forward, the future of generative AI lies in creatively chaining all sorts of LLMs and knowledge bases together to create new kinds of assistants that deliver authoritative results users can verify.

Get a hands on using retrieval-augmented generation with an AI chatbot in this NVIDIA LaunchPad lab.

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 | TensorRT | Trustworthy AI



All NVIDIA News

NVIDIA AI Microservices for Drug Discovery, Digital Health Now Integrated With AWS

NVIDIA Leaders Honored for Outstanding Achievements by Silicon Valley YWCA

GeForce NOW Delivers 24 A-May-zing Games This Month

Explainable AI: Insights from Arthur's Adam Wenchel

AI Takes a Bow: Interactive GLaDOS Robot Among 9 Winners in Hackster.io Challenge

Igniting the Future: TensorRT-LLM Release Accelerates AI Inference Performance, Adds Support for New Models Running on RTX-Powered Windows 11 PCs

New tools and resources announced at Microsoft Ignite include TensorRT-LLM wrapper for OpenAI Chat API, RTX-powered performance improvements to DirectML for Llama 2, other popular LLMs.

November 15, 2023 by Jesse Clayton

Share

Reading Time: 4 mins

Artificial intelligence on Windows 11 PCs marks a pivotal moment in tech history, revolutionizing experiences for gamers, creators, streamers, office workers, students and even casual PC users.

It offers unprecedented opportunities to enhance productivity for users of the more than 100 million Windows PCs and workstations that are powered by RTX GPUs. And NVIDIA RTX technology is making it even easier for developers to create AI applications to change the way people use computers.

New optimizations, models and resources announced at Microsoft Ignite will help developers deliver new end-user experiences, quicker.

An upcoming update to TensorRT-LLM — open-source software that increases AI inference performance — will add support for new large language models and make demanding AI workloads more accessible on desktops and laptops with RTX GPUs starting at 8GB of VRAM.

TensorRT-LLM for Windows will soon be compatible with OpenAI's popular Chat API through a new wrapper. This will enable hundreds of developer projects and applications to run locally on a PC with RTX, instead of in the cloud — so users can keep private and proprietary data on Windows 11 PCs.

Custom generative AI requires time and energy to maintain projects. The process can become incredibly complex and time-consuming, especially when trying to collaborate and deploy across multiple environments and platforms.

AI Workbench is a unified, easy-to-use toolkit that allows developers to quickly create, test and customize pretrained generative AI models and LLMs on a PC or workstation. It provides developers a single platform to organize their AI projects and tune models to specific use cases.

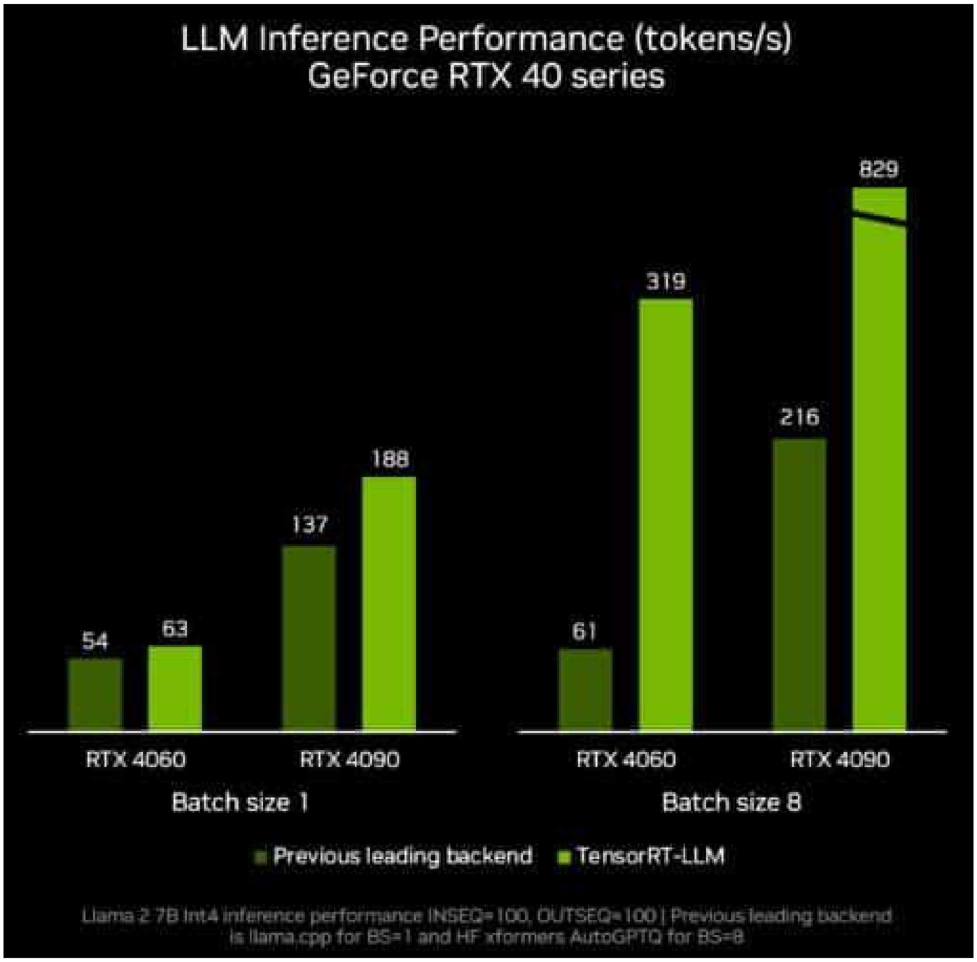
This enables seamless collaboration and deployment for developers to create cost-effective, scalable generative AI models quickly. Join the early access list to be among the first to gain access to this growing initiative and to receive future updates.

To support AI developers, NVIDIA and Microsoft will release DirectML enhancements to accelerate one of the most popular foundational AI models, Llama 2. Developers now have more options for cross-vendor deployment, in addition to setting a new standard for performance.

Portable AI

Last month, NVIDIA announced TensorRT-LLM for Windows, a library for accelerating LLM inference.

The next TensorRT-LLM release, v0.6.0 coming later this month, will bring improved inference performance — up to 5x faster — and enable support for additional popular LLMs, including the new Mistral 7B and Nemotron-3 8B. Versions of these LLMs will run on any GeForce RTX 30 Series and 40 Series GPU with 8GB of RAM or more, making fast, accurate, local LLM capabilities accessible even in some of the most portable Windows devices.



Up to 5X performance with the new TensorRT-LLM v0.6.0.

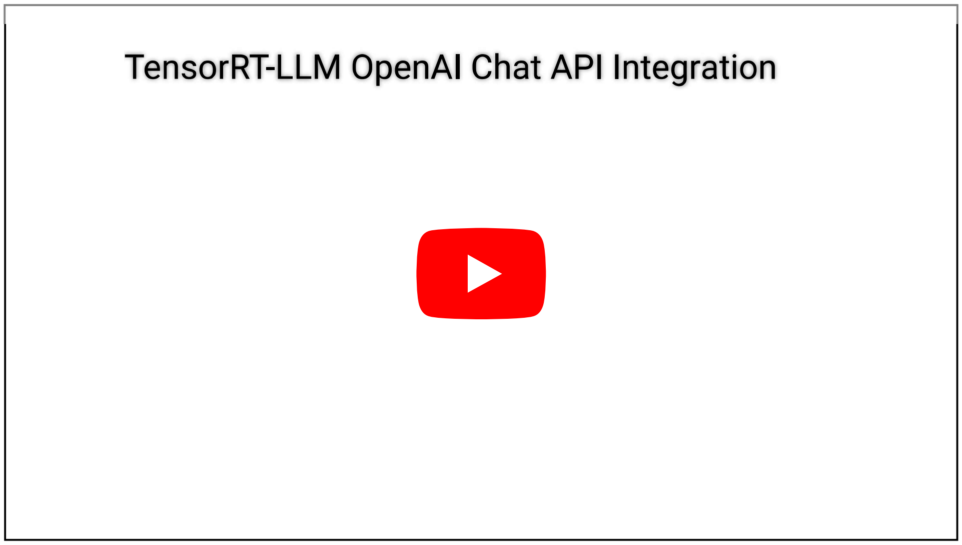
The new release of TensorRT-LLM will be available for install on the /NVIDIA/TensorRT-LLM GitHub repo. New optimized models will be available on ngc.nvidia.com.

Conversing With Confidence

Developers and enthusiasts worldwide use OpenAI's Chat API for a wide range of applications — from summarizing web content and drafting documents and emails to analyzing and visualizing data and creating presentations.

One challenge with such cloud-based AIs is that they require users to upload their input data, making them impractical for private or proprietary data or for working with large datasets.

To address this challenge, NVIDIA is soon enabling TensorRT-LLM for Windows to offer a similar API interface to OpenAI's widely popular ChatAPI, through a new wrapper, offering a similar workflow to developers whether they are designing models and applications to run locally on a PC with RTX or in the cloud. By changing just one or two lines of code, hundreds of AI-powered developer projects and applications can now benefit from fast, local AI. Users can keep their data on their PCs and not worry about uploading datasets to the cloud.



Perhaps the best part is that many of these projects and applications are open source, making it easy for developers to leverage and extend their capabilities to fuel the adoption of generative AI on Windows, powered by RTX.

The wrapper will work with any LLM that's been optimized for TensorRT-LLM (for example, Llama 2, Mistral and Nemotron-3 8B) and is being released as a reference project on GitHub, alongside other developer resources for working with LLMs on RTX.

Model Acceleration

Developers can now leverage cutting-edge AI models and deploy with a cross-vendor API. As part of an ongoing commitment to empower developers, NVIDIA and Microsoft have been working together to accelerate Llama on RTX via the DirectML API.

Building on the announcements for the fastest inference performance for these models announced last month, this new option for cross-vendor deployment makes it easier than ever to bring AI capabilities to PC.

Developers and enthusiasts can experience the latest optimizations by downloading the latest ONNX runtime and following the installation instructions from Microsoft, and installing the latest driver from NVIDIA, which will be available on Nov. 21.

These new optimizations, models and resources will accelerate the development and deployment of AI features and applications to the 100 million RTX PCs worldwide, joining the more than 400 AI-powered apps and games already accelerated by RTX GPUs.

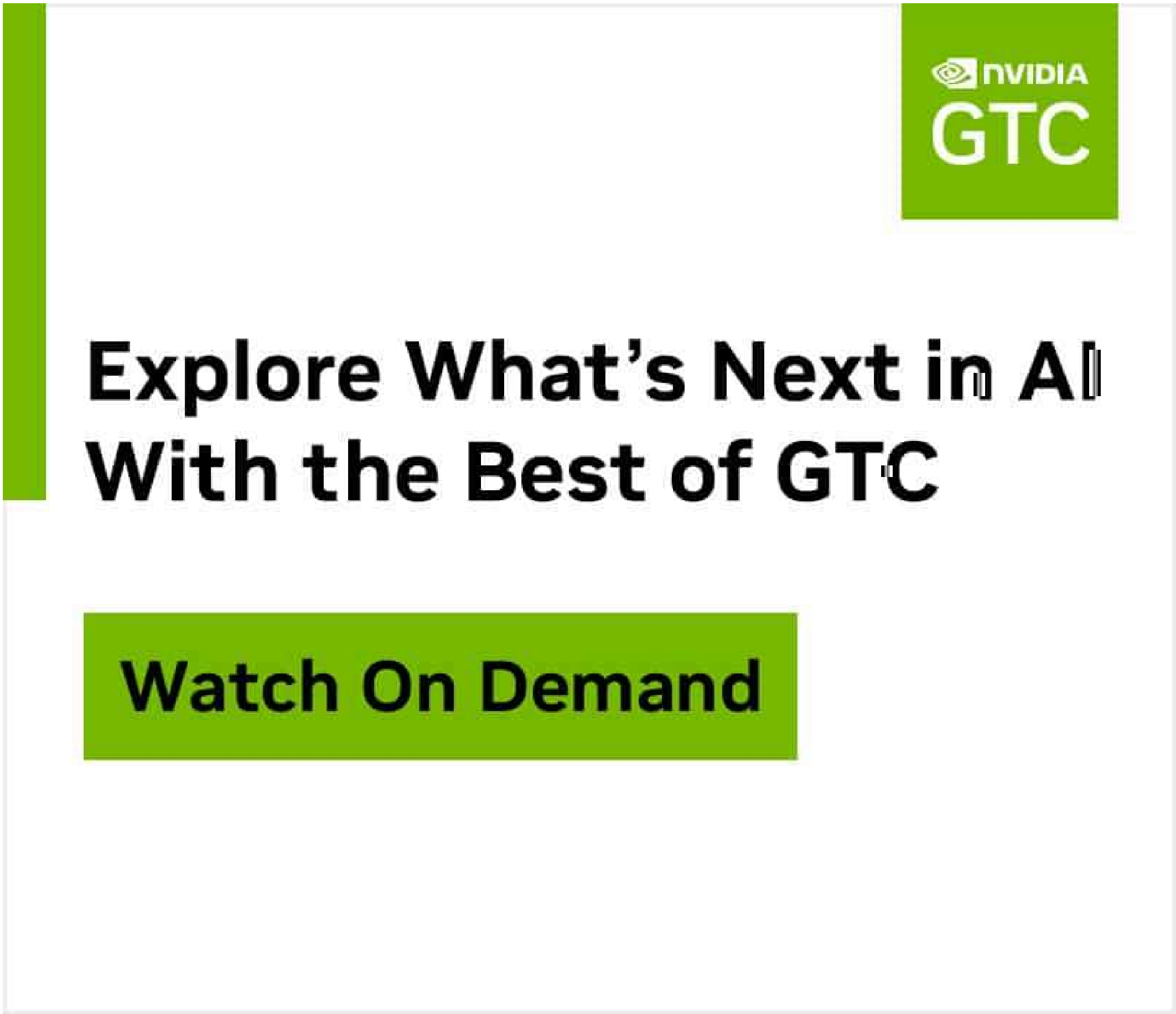
As models become even more accessible and developers bring more generative AI-powered functionality to RTX-powered Windows PCs, RTX GPUs will be critical for enabling users to take advantage of this powerful technology.

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 | Generative AI

Tags: Artificial Intelligence | GeForce | NVIDIA RTX | TensorRT

Load Comments



All NVIDIA News

NVIDIA AI Microservices for Drug Discovery, Digital Health Now Integrated With AWS

NVIDIA Leaders Honored for Outstanding Achievements by Silicon Valley YWCA

GeForce NOW Delivers 24 A-May-zing Games This Month

Explainable AI: Insights from Arthur's Adam Wenchel

AI Takes a Bow: Interactive GLaDOS Robot Among 9 Winners in Hackster.io Challenge

Challenge Accepted: Animator Sir Wade Neistadt Leads Robotic Revolution in Record Time This Week 'In the NVIDIA Studio'

The Razer Blade 18 laptop powered by GeForce RTX 4090 graphics elevated his creative workflow.

November 14, 2023 by Gerardo Delgado

Share

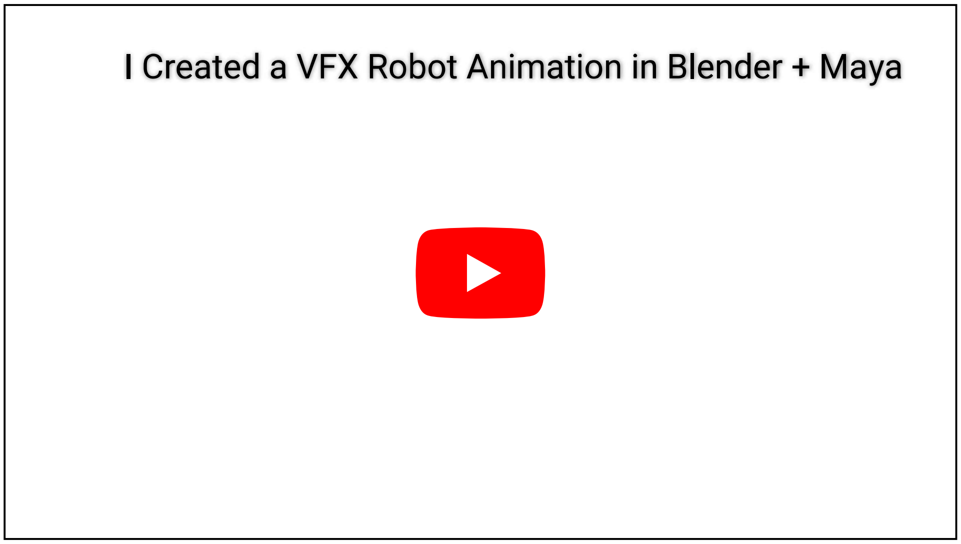
Reading Time: 4 mins

Editor's note: This post is part of our weekly In the NVIDIA Studio series, which celebrates featured artists, offers creative tips and tricks, and demonstrates how NVIDIA Studio technology improves creative workflows. We're also deep diving on new GeForce RTX 40 Series GPU features, technologies and resources, and how they dramatically accelerate content creation.

Character animator Sir Wade Neistadt works to make animation and 3D education more accessible for aspiring and professional artists alike through video tutorials and industry training.

The YouTube creator, who goes by Sir Wade, also likes a challenge. When electronics company Razer recently asked him to create something unique and creative using the new Razer Blade 18 laptop with GeForce RTX 4090 graphics, Sir Wade obliged.

"I said yes because I thought it'd be a great opportunity to try something creatively risky and make something I didn't yet know how to achieve," the artist said.



I Created a VFX Robot Animation in Blender + Maya

I, Robot

One of the hardest parts of getting started on a project is needing to be creative on demand, said Sir Wade. For the Razer piece, the animator started by asking himself two questions: "What am I inspired by?" and "What do I have to work with?"

Sir Wade finds inspiration in games, technology, movies, people-watching and conversations. Fond of tech — and having eyed characters from the ProRigs library for some time — he decided his short animation should feature robots.

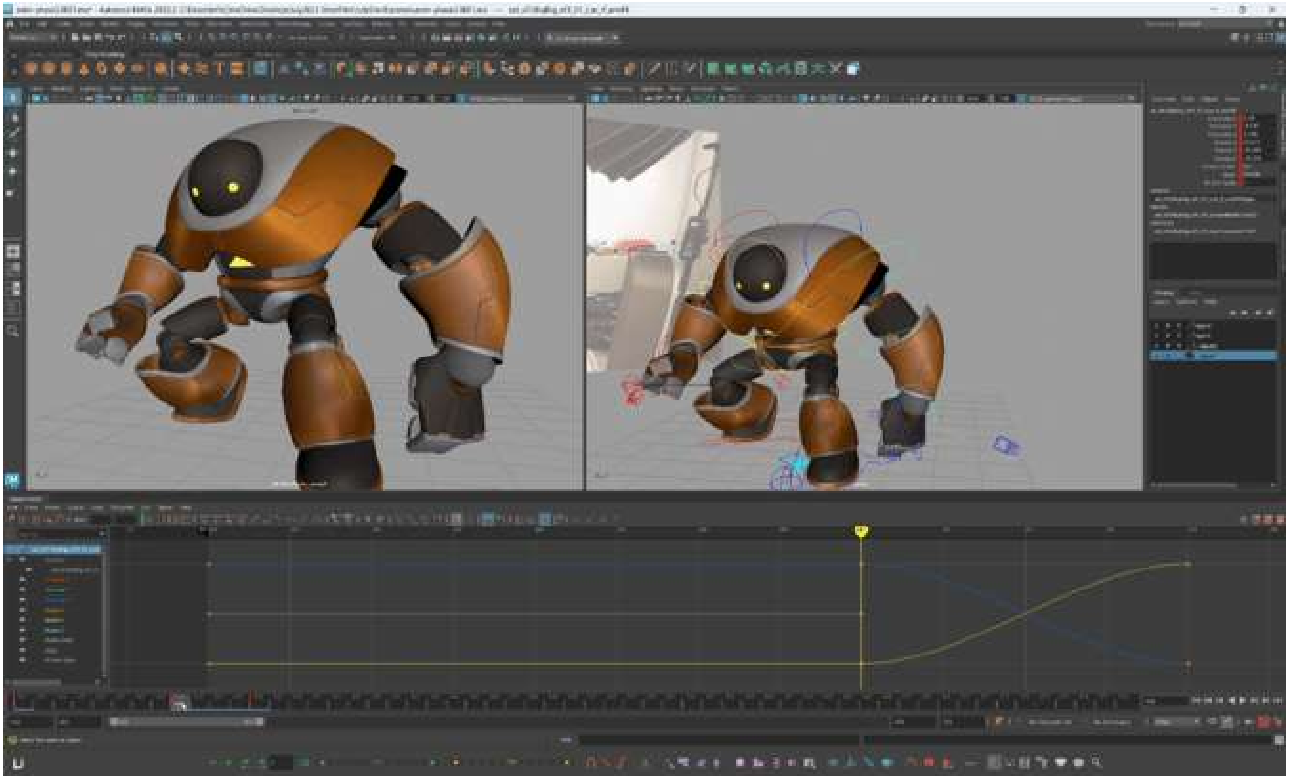
When creating a concept for the animation, Sir Wade took an unorthodox approach, skipping the popular step of 2D sketching. Instead, he captured video references by acting out the animations himself.

This gave Sir Wade the opportunity to quickly try a bunch of movements and preview body mechanics for the animation phase. Since ProRigs characters are rigs based on Autodesk Maya, he naturally began his animation work using this 3D software.

"YOU SHALL NOT (RENDER) PASS."

His initial approach was straightforward: mimicking the main robot character's movements with the edited reference footage. This worked fairly well, as NVIDIA RTX-accelerated ray tracing and AI denoising with the default Autodesk Arnold renderer resulted in smooth viewport movement and photorealistic visuals.

Then, Sir Wade continued tinkering with the piece, focusing on how the robot's arm plates crashed into each other and how its feet moved. This was a great challenge, but he kept moving on the project. The featured artist would advise, "Don't wait for everything to be perfect."



The video reference footage captured earlier paid off later in Sir Wade's creative workflow.

Next, Sir Wade exported files into Blender software with the Universal Scene Description (OpenUSD) framework, unlocking an open and extensible ecosystem, including the ability to make edits in NVIDIA Omniverse, a development platform for building and connecting 3D tools and applications. The edits could then be captured in the original native files, eliminating the need for tedious uploading, downloading and file reformatting.

AI-powered RTX-accelerated OptiX ray tracing in the viewport allowed Sir Wade to manipulate the scene with ease.

Sir Wade browsed the Kitbash3D digital platform with the new asset browser Cargo to compile kits, models and materials, and drag them into Blender with ease. It's important at this stage to get base-level models in the scene, he said, so the environment can be further refined.

Dubbed the "ultimate desktop replacement," the Razer Blade 18 offers NVIDIA GeForce RTX 4090 graphics.

Sir Wade raved about the Razer Blade 18's quad-high-definition (QHD+) 18″ screen and 16:10 aspect ratio, which gives him more room to create, as well as its color-calibrated display, which ensures uploads to social media are as accurate as possible and require minimal color correction.

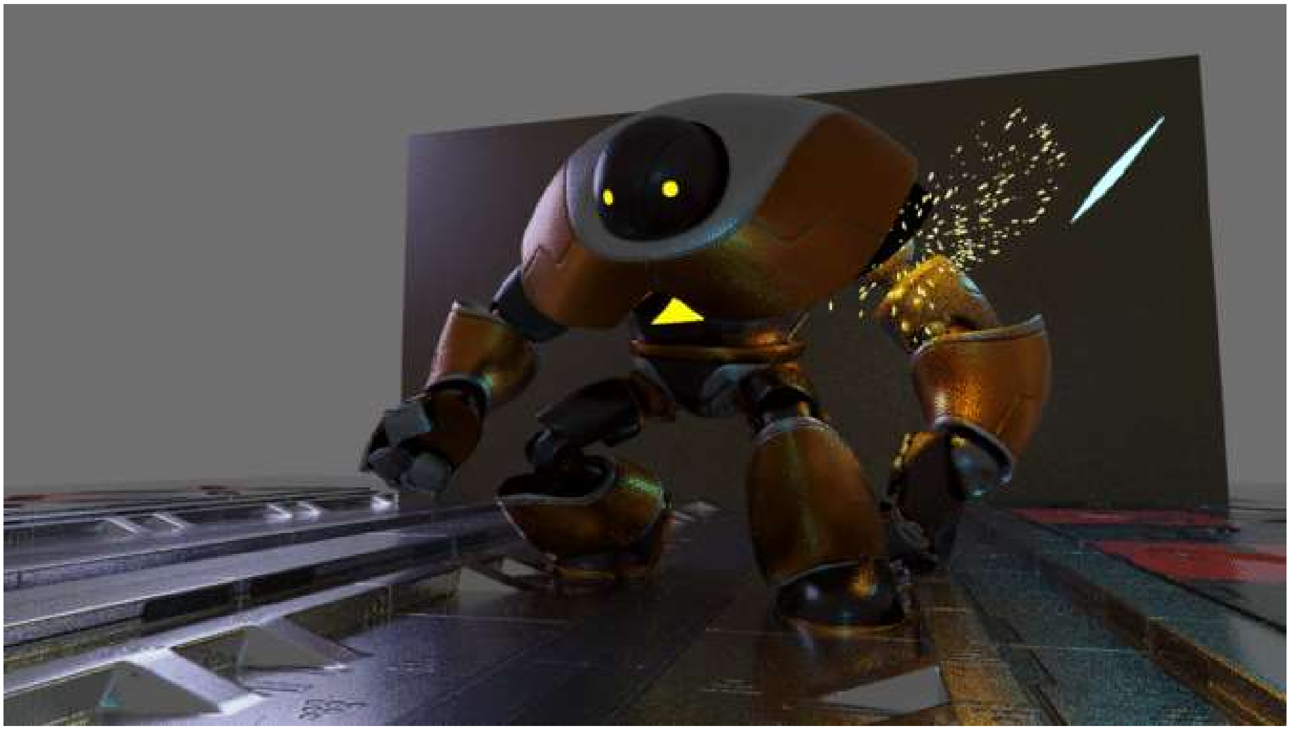
The preinstalled NVIDIA Studio Drivers, free to RTX GPU owners, are extensively tested with the most popular creative software to deliver maximum stability and performance.

Returning to the action, Sir Wade used an emission shader to form the projectiles aimed at the robot. He also tweaked various textures, such as surface imperfections, to make the robot feel more weathered and battle-worn, before moving on to visual effects (VFX).

> "This is by far the best laptop I've ever used for this type of work." — Sir Wade Neistadt

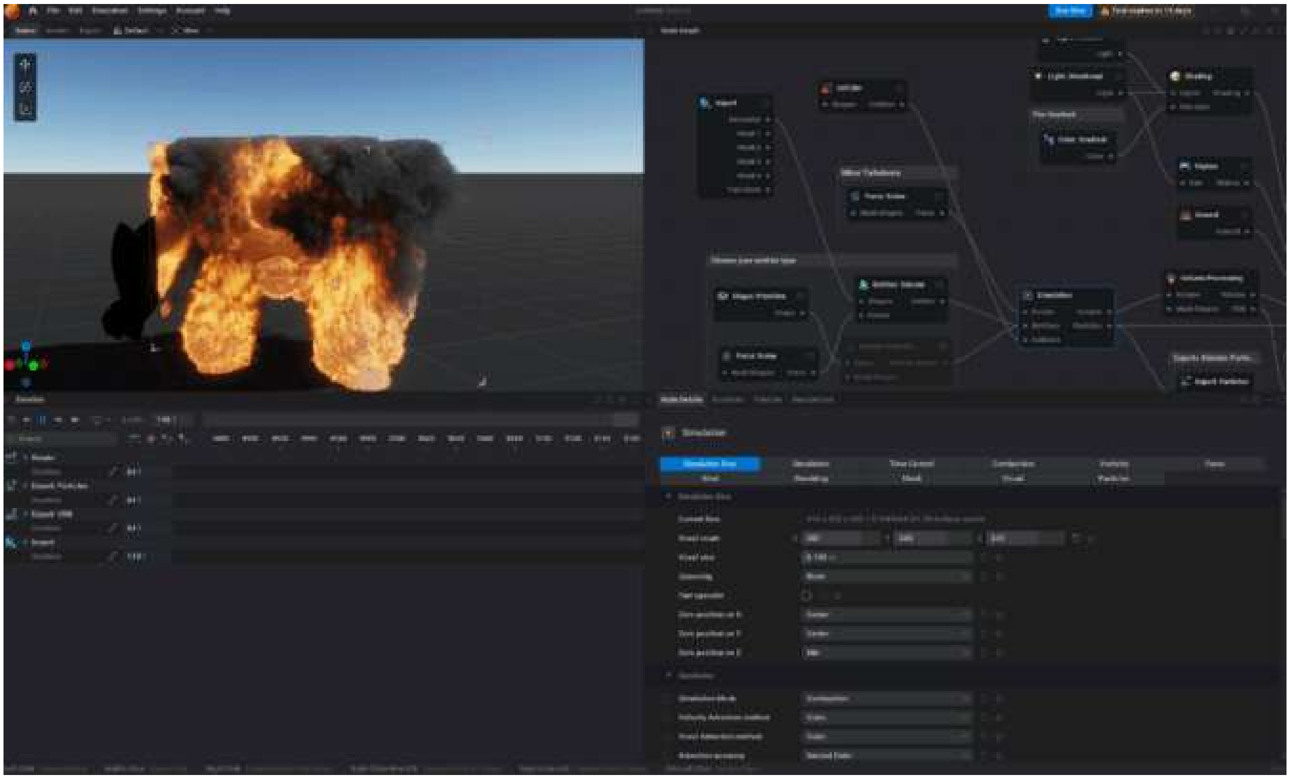
The artist used basic primitives as particle emitters in Blender to achieve the look of bursting particles over a limited number of frames. This, combined with the robot and floor surfaces containing surface nodes, creates sparks when the robot moves or gets hit by objects.

Sir Wade's GeForce RTX 4090 Laptop GPU with Blender Cycles RTX-accelerated OptiX ray tracing in the viewport provides interactive, photorealistic rendering for modeling and animation.



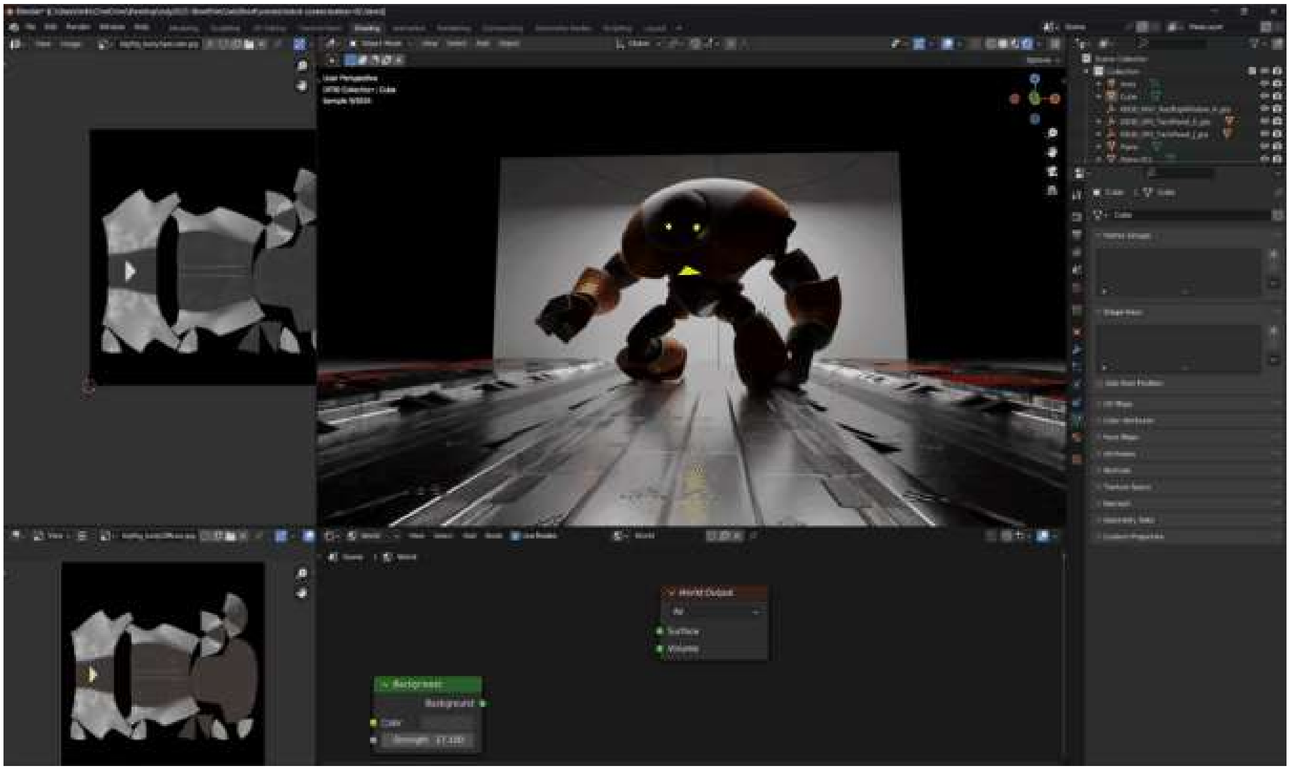
Particle and collusion effects in Blender enable compelling VFX.

To further experiment with VFX, Sir Wade imported the project into the EmberGen simulation tool to test out various preset and physics effects.



VFX in EmberGen.

He added dust and debris VFX, and exported the scene as an OpenVDB file back to Blender to perfect the lighting.



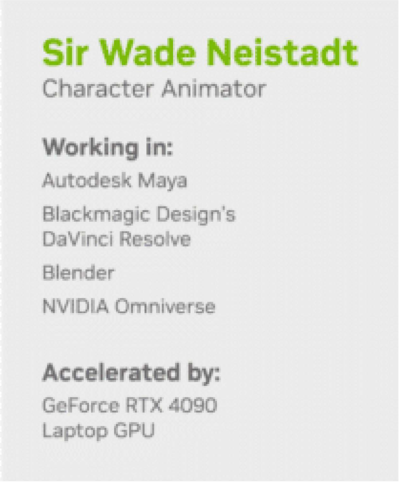
Final lighting elements in Blender.

Finally, Sir Wade completed sound-design effects in Blackmagic Design's DaVinci Resolve software.

Sir Wade's video tutorials resonate with diverse audiences because of their fresh approach to solving problems and individualistic flair.

"Creativity for me doesn't come naturally like for other artists," Sir Wade explained. "I reverse engineer the process by seeing a tool or a concept, evaluating what's interesting, then either figuring out a way to use it uniquely or explaining the discovery in a relatable way."

> "I chose an NVIDIA RTX GPU-powered system for its reliable speed, performance and stability, as I had a very limited window to complete this project." — Sir Wade Neistadt



Sir Wade Neistadt
Character Animator

Working in:
Autodesk Maya
Blackmagic Design's DaVinci Resolve
Blender
NVIDIA Omniverse

Accelerated by:
GeForce RTX 4090 Laptop GPU

Sir Wade Neistadt.

Check out Sir Wade's animation workshops on his website.

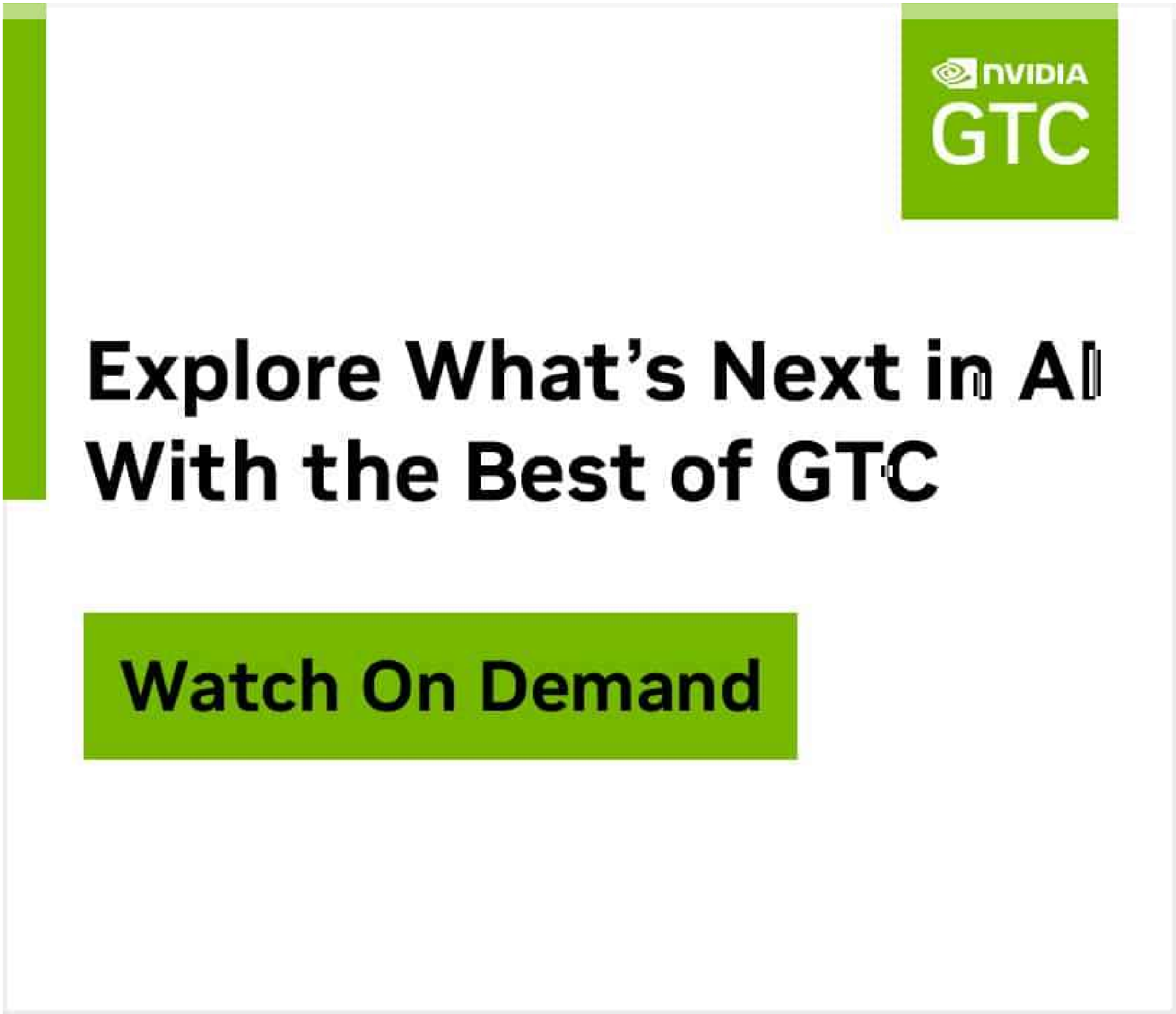
Less than two days remain in Sir Wade's Fall 2023 Animation Challenge. Download the challenge template and Maya character rig files, and submit a custom 3D scene to win an NVIDIA RTX GPU or other prizes by end of day on Wednesday, Nov. 15.

Follow NVIDIA Studio on Instagram, Twitter and Facebook. Access tutorials on the Studio YouTube channel and get updates directly in your inbox by subscribing to the Studio newsletter.

Categories: Pro Graphics

Tags: 3D | Art | Artificial Intelligence | Creators | GeForce | In the NVIDIA Studio | NVIDIA Studio | Rendering

Load Comments



All NVIDIA News

NVIDIA AI Microservices for Drug Discovery, Digital Health Now Integrated With AWS

NVIDIA Leaders Honored for Outstanding Achievements by Silicon Valley YWCA

GeForce NOW Delivers 24 A-May-zing Games This Month

Explainable AI: Insights from Arthur's Adam Wenchel

AI Takes a Bow: Interactive GLaDOS Robot Among 9 Winners in Hackster.io Challenge