



A Chatbot is born - or how to use the magic of GenAI with NVIDIA AI Enterprise

Michael Gern
Florian Schorn



Authorized Distributor





Michael Gern

Solutions Architect Data & AI

Michael Gern is a solutions architect at TD SYNnex EMEA .

He designs and validates partner solutions with a focus on AI and in particular NVIDIA

Supporting the local teams and their resellers with Training, Enablement and pre-sales support.

Besides his main job he drives cross Business-Unit activities to scale the AI message across TD SYNnex`s ecosystem.

TD SYNnex Europe GmbH

Kistlerhofstraße 75

81379 Munich

E-Mail: michael.gern@tdsynnex.com



LinkedIn



Florian Schorn

Supervisor Technical PreSales NVIDIA AI & Nutanix

Florian Schorn Supervisor Presales Consultant at TD SYNnex Germany, specializing in NVIDIA AI & Nutanix solutions. He provides comprehensive support to clients, including training, webinars, consulting, sizing, partner certification, and installation services. Florian is a certified Nutanix instructor and systems engineer, enabling him to deliver in-depth technical guidance.

TD SYNnex Germany GmbH & Co. OHG
Kistlerhofstraße 75
81379 Munich

E-Mail: florian.schorn@tdsynnex.com
Tel.: +49 (0) 151 / 727 001 58



[LinkedIn](#)

Please create me a tour in Barcelona. Starting and Ending at the Placa Espania. Around 15 km with the most popular signseeing attractions



Here's a 15 km walking tour in Barcelona that starts and ends at Plaça d'Espanya, covering some of the city's most popular attractions:

1. Start at Plaça d'Espanya

- The tour begins at Plaça d'Espanya, one of Barcelona's major squares, known for its iconic

Can you please build me a file that i can import in Komoot



Here is the GPX file for your Barcelona walking tour: [Download Barcelona Walking Tour GPX](#). You can import it into Komoot to follow the route. [-]



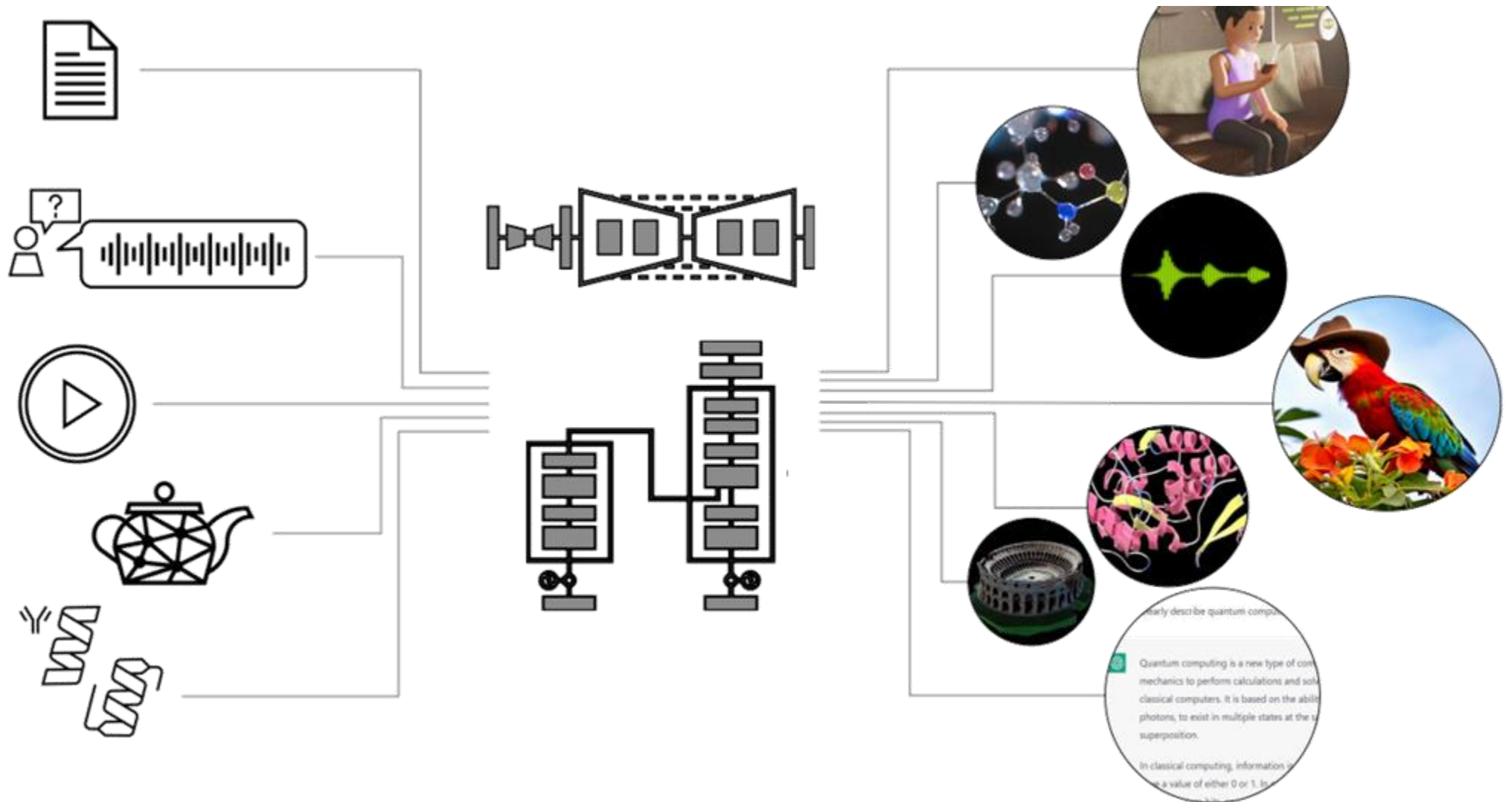
- Descend towards the **Barceloneta neighborhood** and head to **Barceloneta Beach**. This lively beach area is perfect for a short break along the Mediterranean coast.
- Walk along the seaside promenade to **Port Olímpic**, a marina area with restaurants and public art installations.

4. La Rambla and Gothic Quarter (Barri Gòtic)

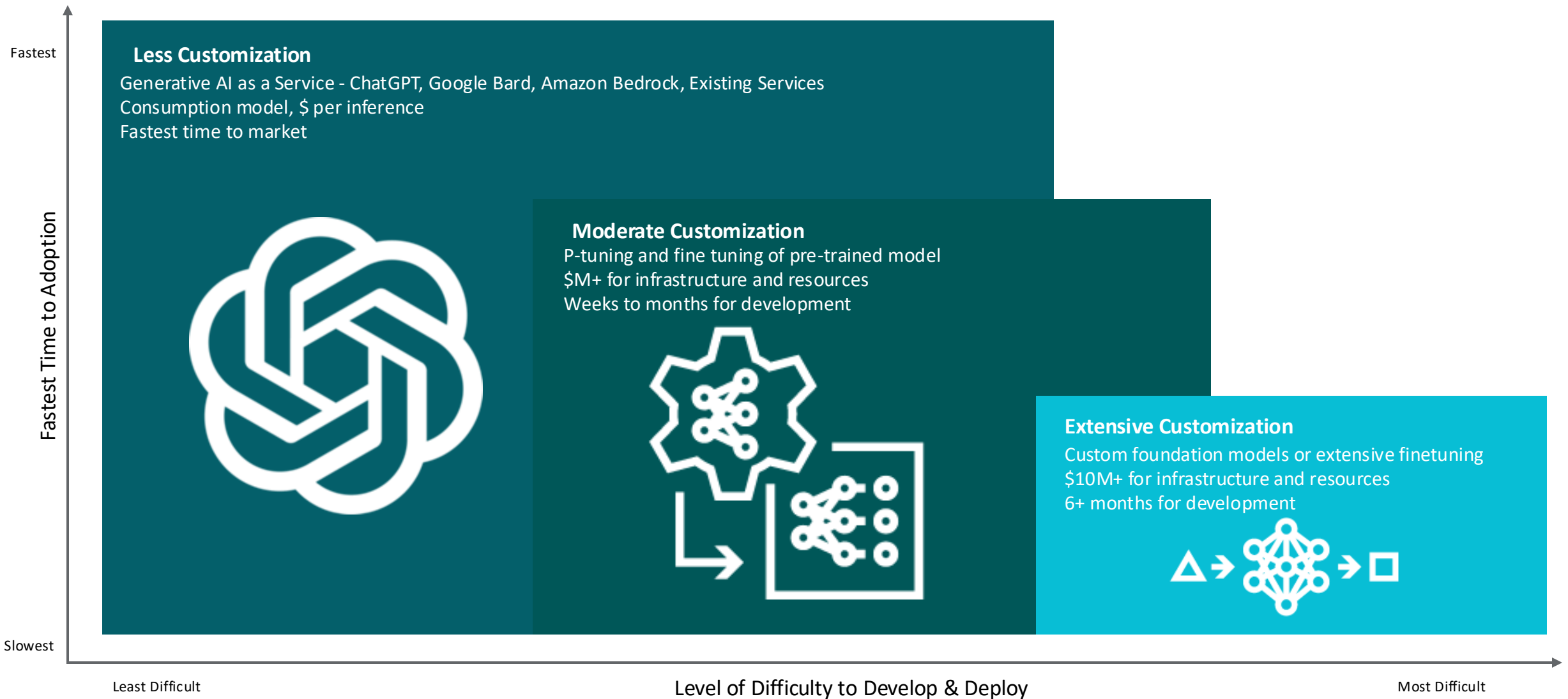
- From the coast, head to **La Rambla**, the famous pedestrian boulevard. Enjoy the bustling atmosphere and maybe stop at the **Mercat de Sant Josep de la Boqueria**, a vibrant food market.



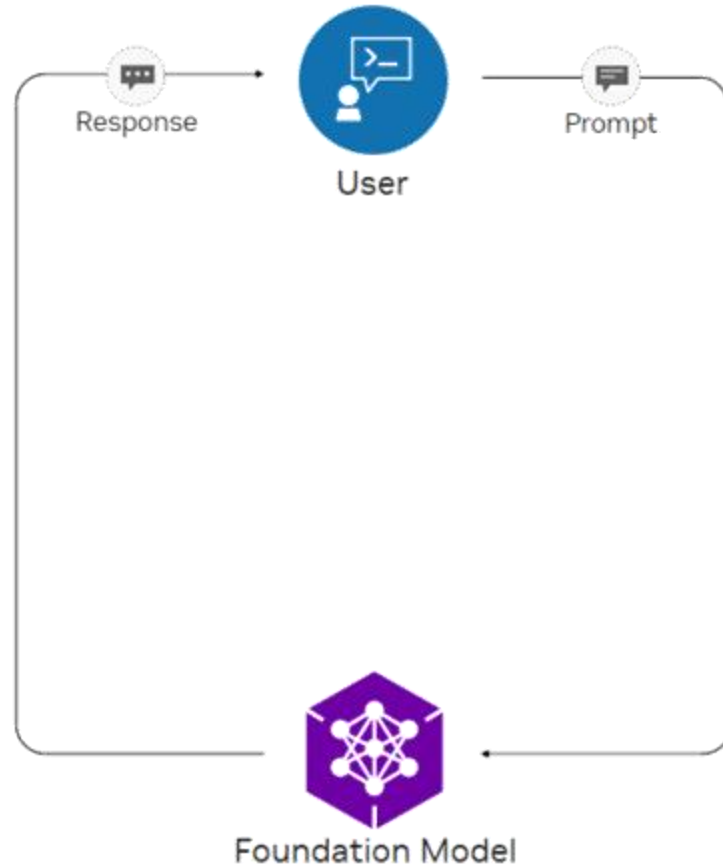
What is Generative AI (Gen AI)



How to consume AI models




Community LLMs are not powerful enough



Lacking proprietary knowledge

who won the EM 2024?

 As of my last update in January 2022, I don't have information on who won the UEFA Euro 2024, as it's in the future from my knowledge cutoff. For the most current information, I recommend checking a reliable sports news website or the official UEFA Euro 2024 website.



Hallucinations



llama3.1:latest ▾ +

Set as default



Hello, Floscho

How can I help you today?

⚡ Suggested

Tell me a fun fact
about the Roman Empire

Prompt



Help me study
vocabulary for a college entrance
exam

Prompt



Overcome procrastination
give me tips

Prompt



Explain opt
if I'm familiar
selling stocks

Prompt

+ Send a Message



LLMs can make mistakes. Verify important information.

Wagenklasse. Wenn du nicht bereit bist, mehr Geld für den Beförderungsort auszugeben, ist die erste Wagenklasse möglicherweise nicht die beste Wahl.

- **Fahrten mit Zugrissen** : Bei einigen Zügen werden die Sitze in der ersten Klasse nur während der Fahrt reserviert. In diesen Fällen kann es passieren, dass du ohne vorherige Reservierung auf einen Platz in der ersten Wagenklasse gehst und dieser dann nicht frei ist. Es gibt jedoch auch Züge mit vollen Reservierungen in der ersten Klasse.
- **Buchbarkeit von Fahrkarten** : In einigen Fällen kannst du eine Fahrkarte für die erste Wagenklasse online buchen, aber es kann passieren, dass diese nicht mehr verfügbar ist, wenn du sie im Zug selbst kaufst. Es ist also immer eine gute Idee, in Voraussicht, online zu reservieren.

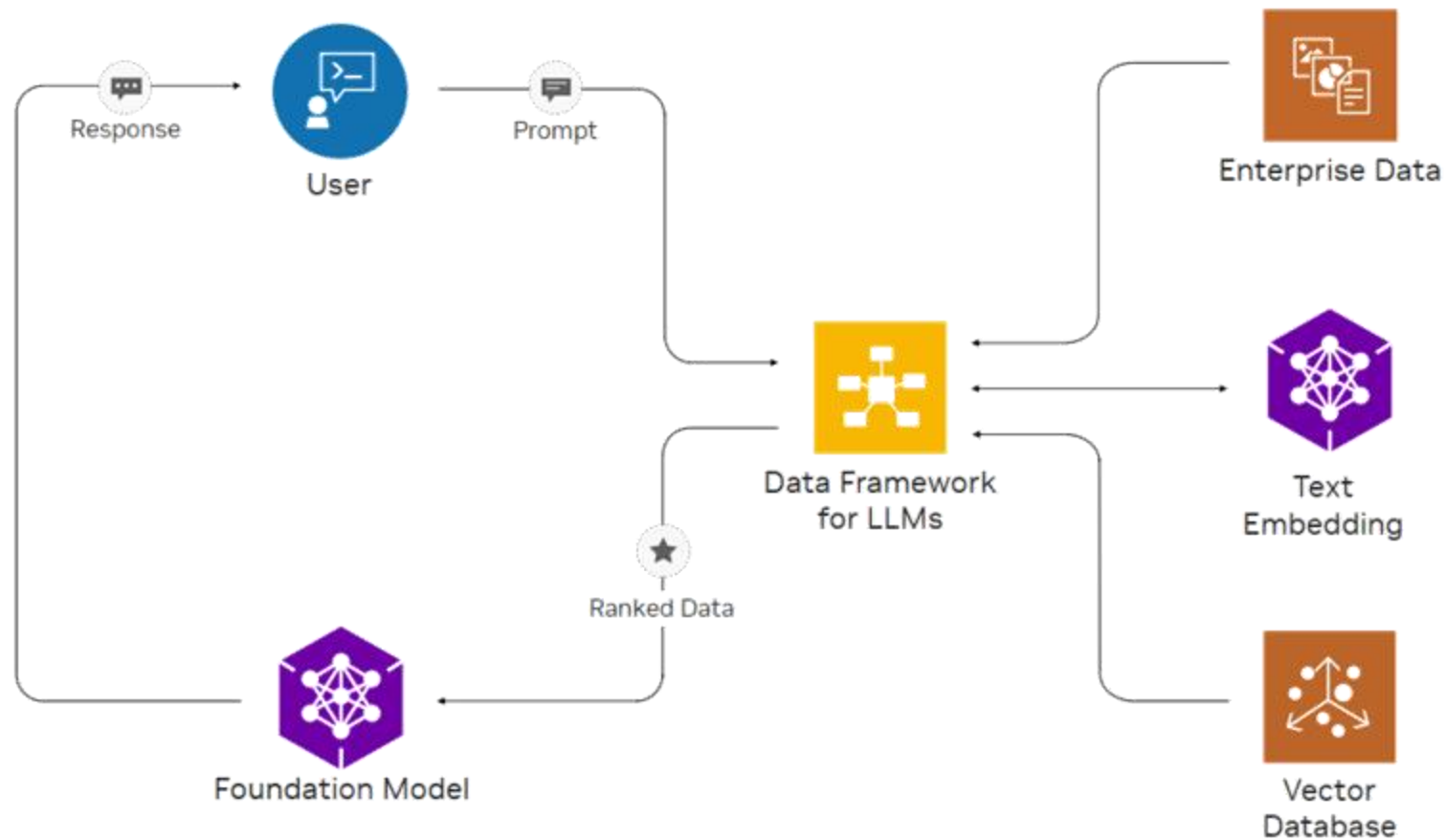
Im Allgemeinen kannst du jedoch auf Bahnreisen 1.Klasse buchen, solange du ein gültiges Fahrkartenprodukt dafür hast.



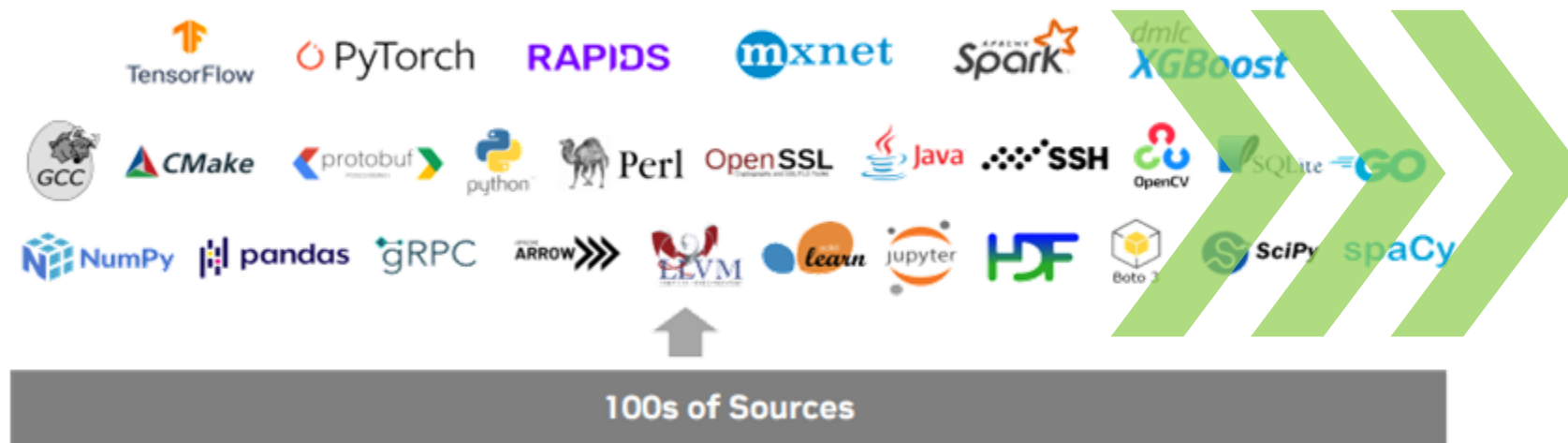
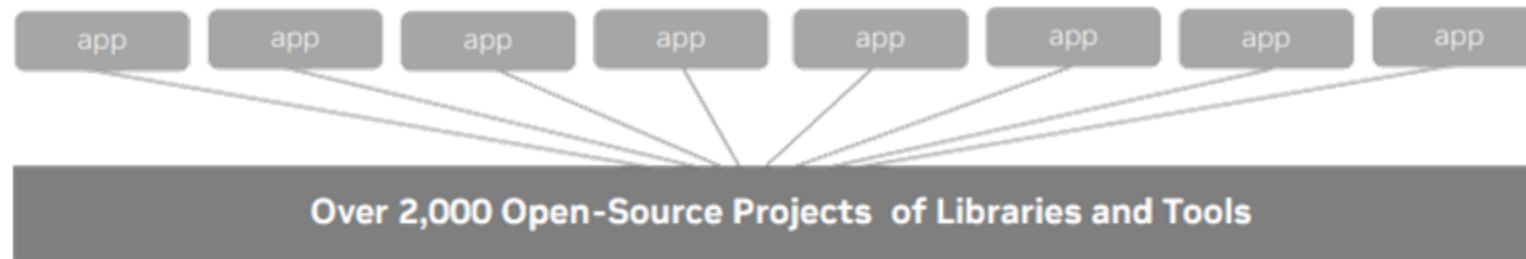
+ Send a Message









RAG – Retrieval Augmented Generation



The Open Source AI stack



 **NVIDIA**. AI Enterprise runs best on  **Red Hat**

	 VMware vSphere*	 Red Hat OpenShift ¹	 Red Hat Enterprise Linux	 kubernetes ²	 Amazon EKS	 Google Kubernetes Engine	HPE E Z M E R A L	Container Only ³
GPU	✓	✓	✓	✓	✓	✓	✓	✓
CPU	✓	✓	✓	✓	✗	✗	✓	✓
Bare-Metal	✗	✓	✓	✓	✗	✗	✓	✓
Virtualized⁴	✓	✓	✓	✓	✗	✗	✓	✓
Public Cloud	✗	✓	✓	✗	✓	✓	✗	✓ ⁵
DGX Systems	✗	✓	✓	✓ ⁶	✗	✗	✗	✓

¹With Red Hat CoreOS

²Upstream Kubernetes only

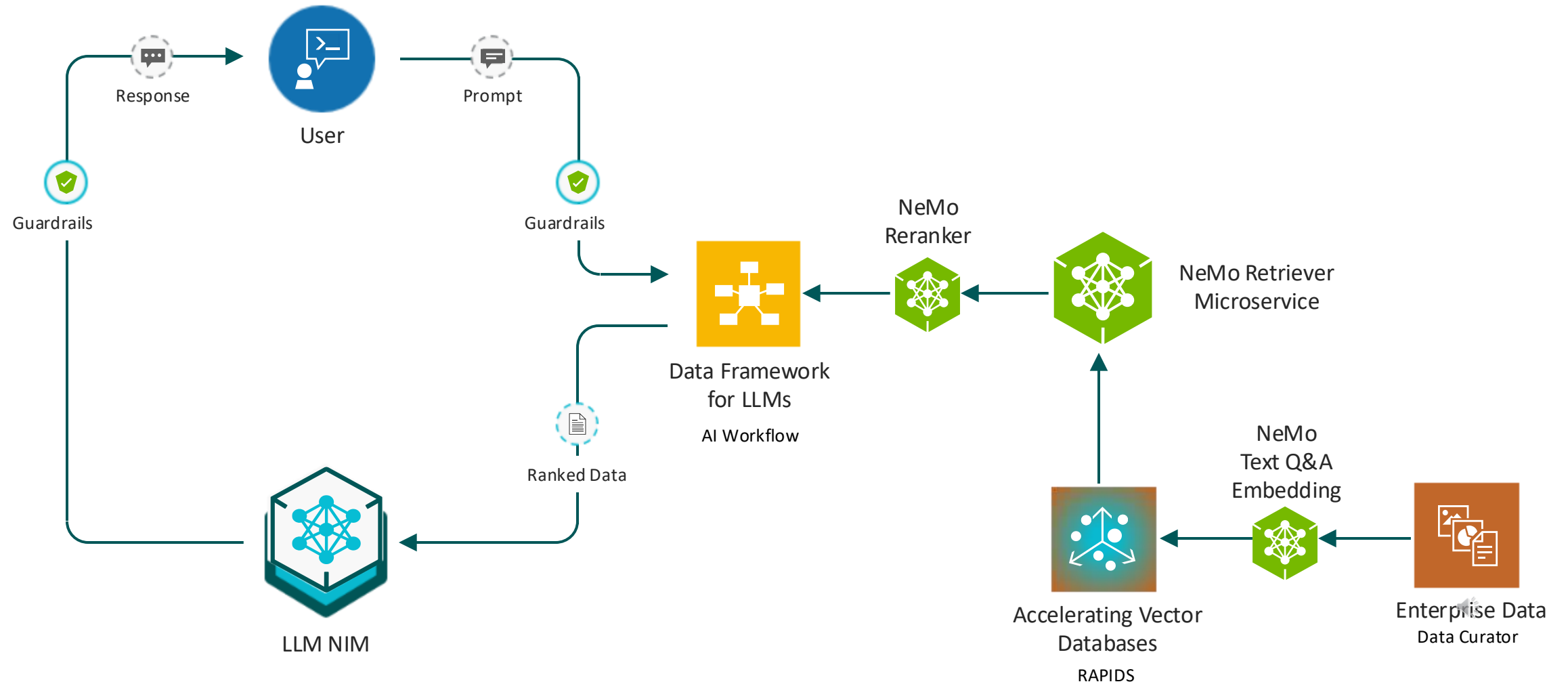
³For RHEL 8.4 and Ubuntu 20.04, no orchestration support needed

⁴With VMware vSphere 7.0 U3c or later or Red Hat Enterprise Linux KVM

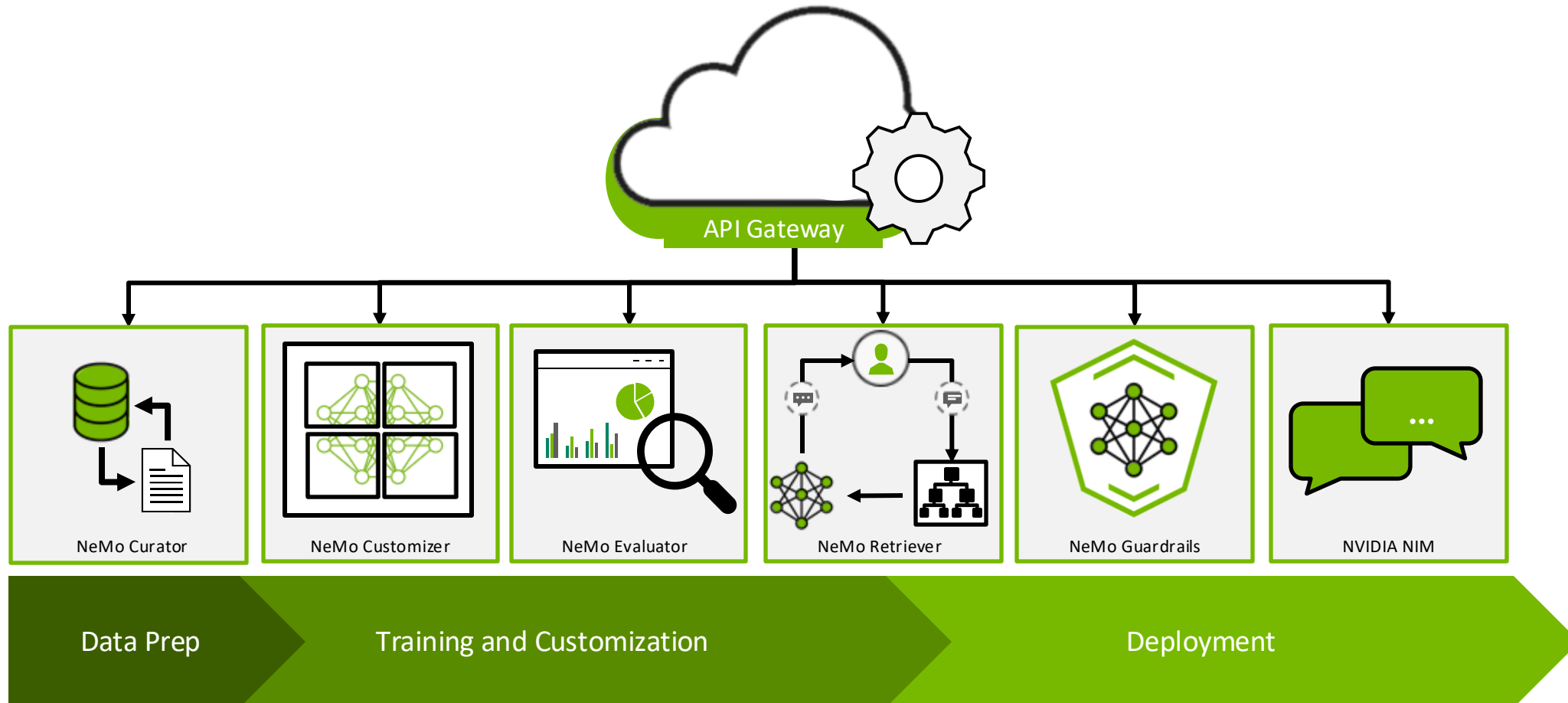
⁵Designated GPU instances on AWS, Google Cloud, Azure

⁶GPU Passthrough only

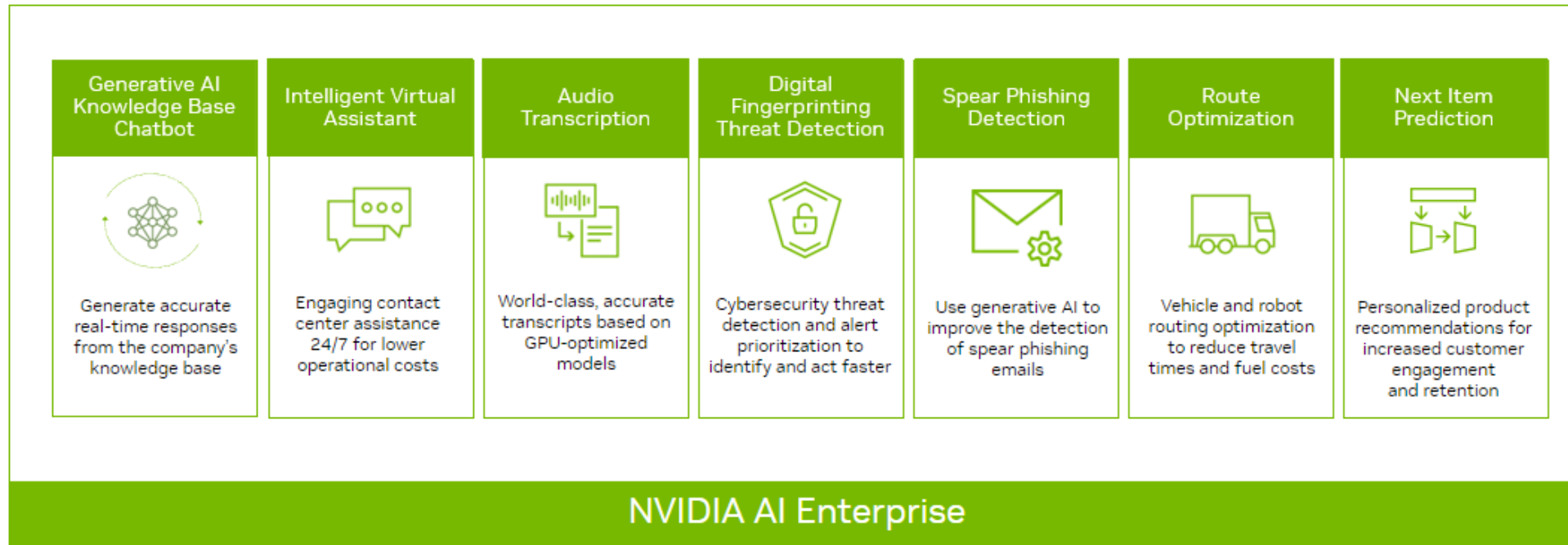
RAG – NVIDIA AI Enterprise Style



Building Generative AI applications for the Enterprise



NVAIE – not only for GenAI



Cloud



Data Center



Edge



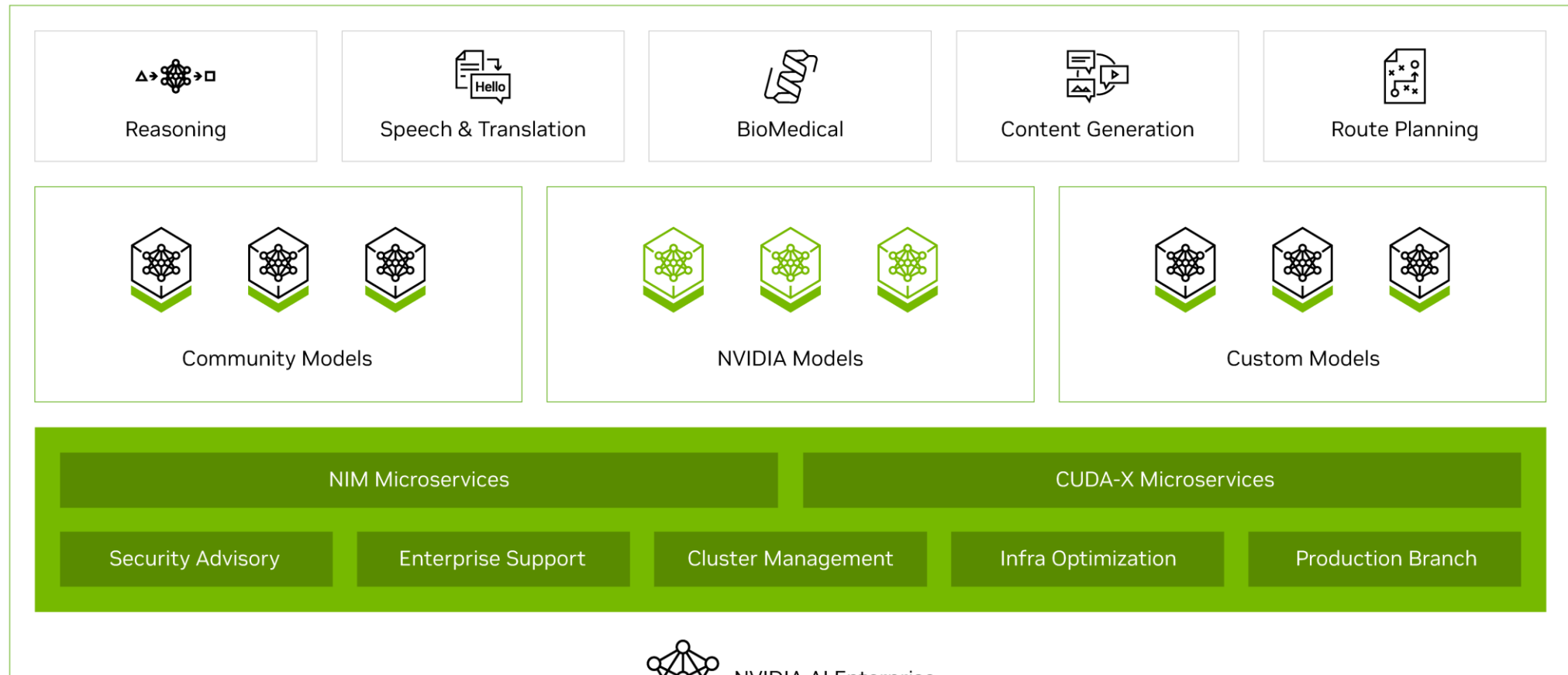
Embedded


runs best on



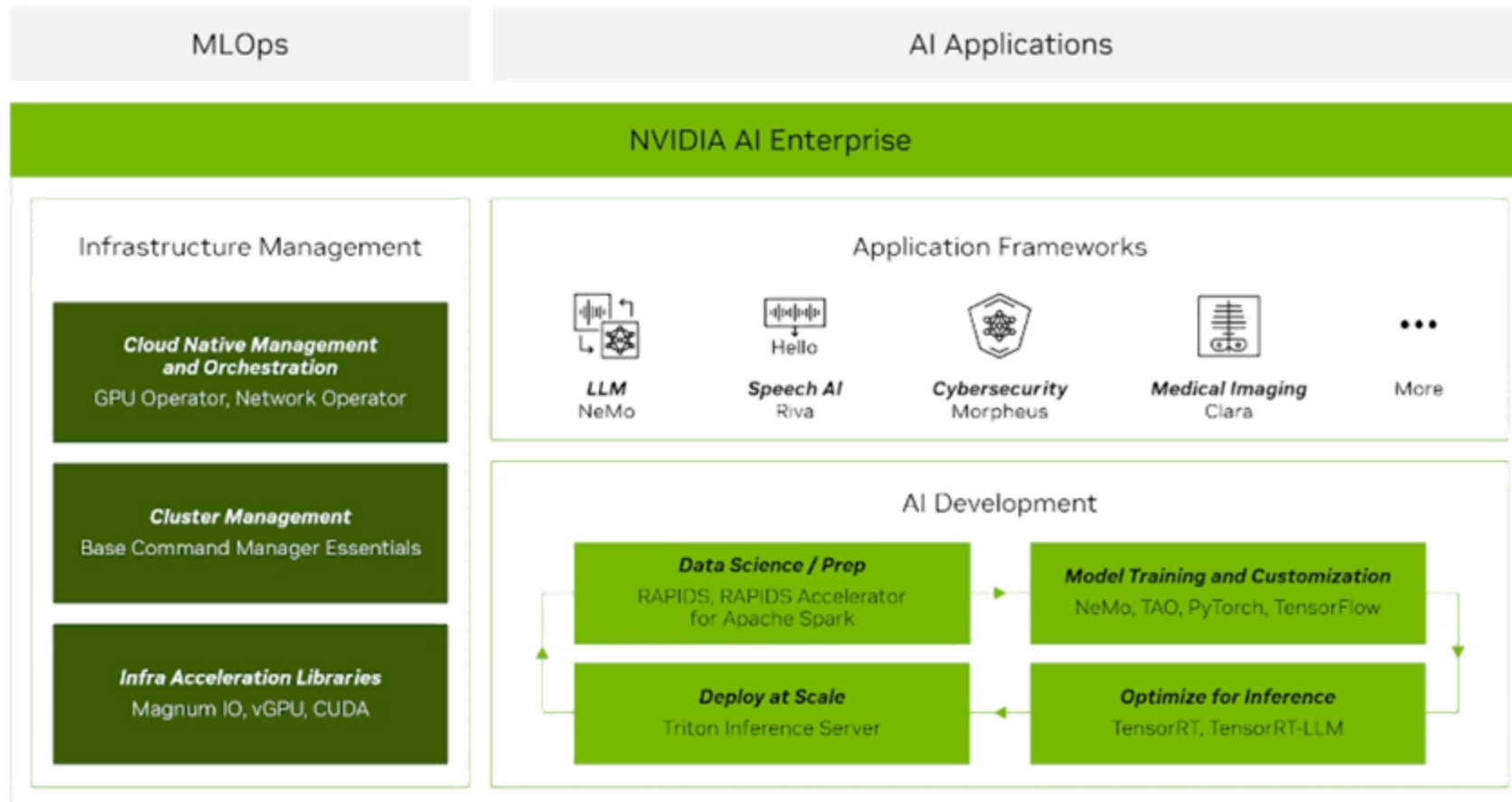
Red Hat

NVAIE – not only for Community Models



runs best on  **Red Hat**

NVAIE – Your Platform for all AI Workloads



runs best on



Red Hat

Usecases



★ Featured

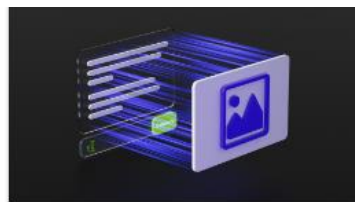
Healthcare & Life Sciences, Biopharma

Biomolecular Generation

Workload: Generative AI / LLMs

Products: NVIDIA BioNeMo

Business Goal: Innovation, Return on Investment



★ Featured

Aerospace, Agriculture, Architecture / Engineering / Constructio...

Accelerating Content Generation

Workload: Generative AI / LLMs

Products: NVIDIA NeMo, NVIDIA Picasso, NVIDIA AI Enterprise

Business Goal: Return on Investment



★ Featured

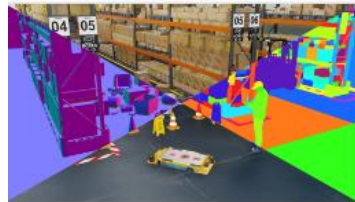
Aerospace, Agriculture, Architecture / Engineering / Constructio...

Digital Human

Workload: Generative AI / LLMs

Products: NVIDIA ACE, NVIDIA Riva, NVIDIA NeMotron, NVIDIA AZX

Business Goal: Innovation



Aerospace, Agriculture, Architecture / Engineering / Constructio...

Synthetic Data Generation

Workload: Computer Vision / Video Analytics

Products: NVIDIA Omniverse Enterprise, NVIDIA DRIVE, NVIDIA Isaac, NVIDIA Metropolis

Business Goal: Innovation



Financial Services

Fraud Detection

Workload: Data Science

Products: NVIDIA AI Enterprise, NVIDIA RAPIDS, NVIDIA Morpheus

Business Goal: Risk Mitigation



Telecommunications

Network Operations

Workload: Generative AI / LLMs

Products: NVIDIA AI Enterprise, NVIDIA Riva, NVIDIA cuOpt, NVIDIA Triton Inference Server, NVIDIA DGX Cloud

Business Goal: Innovation



Telecommunications, Financial Services, Retail/ Consumer Packa...

Agent Assist

Workload: Conversational AI / NLP

Products: NVIDIA AI Enterprise, NVIDIA Riva, NVIDIA DGX

Business Goal: Innovation, Return on Investment



Manufacturing

Virtual Factory

Workload: Simulation / Modeling / Design

Products: NVIDIA Omniverse Enterprise

Business Goal: Innovation



Financial Services

Document Intelligence

Workload: Generative AI / LLMs

Products: NVIDIA DGX, NVIDIA AI Enterprise, NVIDIA NIM, NVIDIA Triton Inference Server, NVIDIA NeMo

Business Goal: Return on Investment, Risk Mitigation

CANCER ADVANCE AI ENTER

Solution Showcase



NVIDIA AI Enterprise Singtel to Unlock New Enterprise Services

"Having a full software stack and rich APIs and libraries is key to supporting the diverse applications and use cases. This accelerated AI Enterprise helps unlock value from data for mission critical use cases"

1 2 3 4 5 6 7 8 9 10 11 12

Solution Brief

Solution Showcase: NVIDIA AI Enterprise
Singtel to Unlock New

Solution Showcase



ControlExpert Revolutionizes Motor Claims Management With NVIDIA AI Enterprise

"We have a vision where drivers around the world can get car damage claims settled fairly in a day. NVIDIA AI Enterprise gave us the performance to provide our customers with real-time responses as well as the security, stability, and support to provide the best customer service 24/7"

Solution Showcase



VNPT Advances Traffic Safety With NVIDIA AI Enterprise

Using the software suite, Vietnam's largest telecom provider increased roadway safety.

"VNPT received excellent support from NVIDIA experts to set up and optimize the AI models, resolve issues, and enhance infrastructure performance in an optimal manner."

Cao Thanh Ha, CTO, VNPT AI

Solution Brief

Solution Showcase: VNTP Advances Traffic Safety with
NVIDIA AI Enterprise



ControlExpert

Customer Profile



VNPT

Customer Profile

Organization: Vietnam Posts and Telecommunications (VNPT)

Case Studys



[Link](#)

All You Need to Build AI. All in C

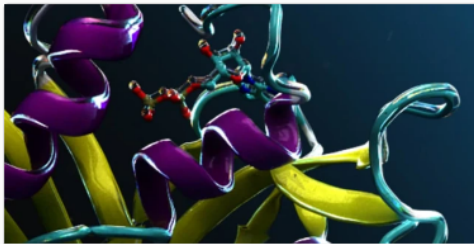
Welcome to the NGC Catalog - GPU Accelerated AI models and SDKs that help you infu

Explore Use Cases



log...

NVIDIA NIM

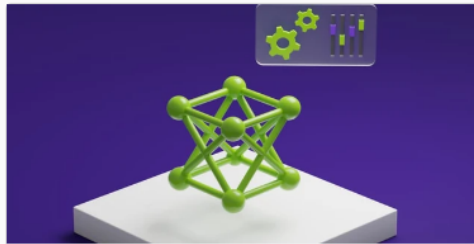


DiffDock
Container

Diffdock predicts the 3D structure of the interaction between a molecule and a protein.

NVIDIA Developer Program +1

NVIDIA NIM +1

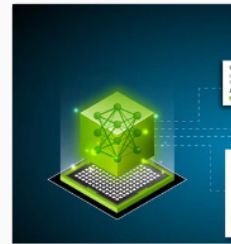


Mixtral-8x7B-Instruct-v0.1
Container

NVIDIA NIM for GPU accelerated Mixtral-8x7B-Instruct-v0.1 inference through OpenAI compatible APIs

NVIDIA Developer Program +1

NVIDIA NIM +1



meta-llama-2-7b
Container

NVIDIA NIM for GPU 7B inference through compatible APIs

NVIDIA Developer Program +1

NVIDIA NIM +1

Getting Started



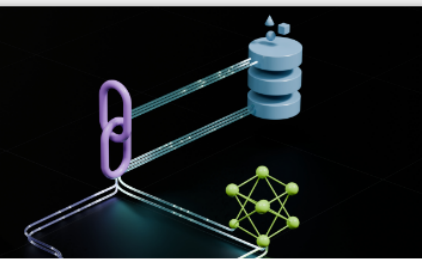
[Link](#)

NVIDIA NGC Catalog

Request Access to Hands-On Labs

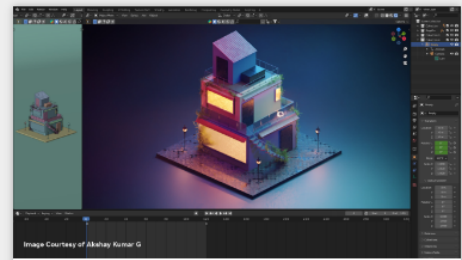
Ready for a hands-on experience with NVIDIA software solutions? Pick a lab to browse

Filters: Industry Technologies Categories Products Use Cases Quick Search



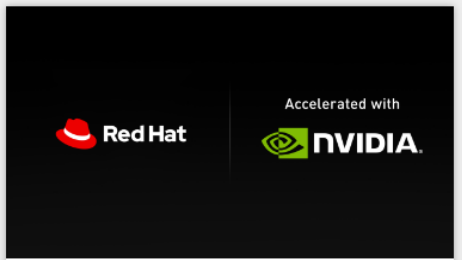
★ FEATURED
Unlock Enterprise Data with NVIDIA NeMo Retriever microservices

Best for: AI practitioner, Data scientist
Included products: Docker, NVIDIA AI Enterprise, NVIDIA GH200 Grace Hopper Superchip
Included technologies: Kubernetes, Langchain, NVIDIA



★ FEATURED
Experience Accelerated Visualization With NVIDIA RTX Virtual Workstation

Best for: Enterprise Graphics Professionals
Included products: NVIDIA RTX Virtual Workstation, NVIDIA-Certified Systems
Included technologies: NVIDIA-



★ FEATURED
Development Framework: Red Hat OpenShift

Best for: AI practitioner, Developer
Included products: NVIDIA L40S GPU

NVIDIA Launchpad



[Link](#)

NVIDIA AI Enterprise

NVIDIA Docs Hub > NVIDIA AI Enterprise

NVIDIA AI Enterprise is an end-to-end, cloud-native software platform that accelerates co-pilots and other generative AI applications. Easy-to-use microservices provide optimal, smooth transition from prototype to production for enterprises that run their businesses.

Overview

Release Documentation

NIMs

AI Workflows

Deployment Guides

AI Chatbot with Retrieval Augmented Generation

Build Generative AI chatbots that accurately answer domain-specific queries using latest information.

[Browse >](#)

Cybersecurity AI Workflows

Reduce the time to develop a solution that can detect cybersecurity threats.

[Browse >](#)

Route Optimization AI Workflow

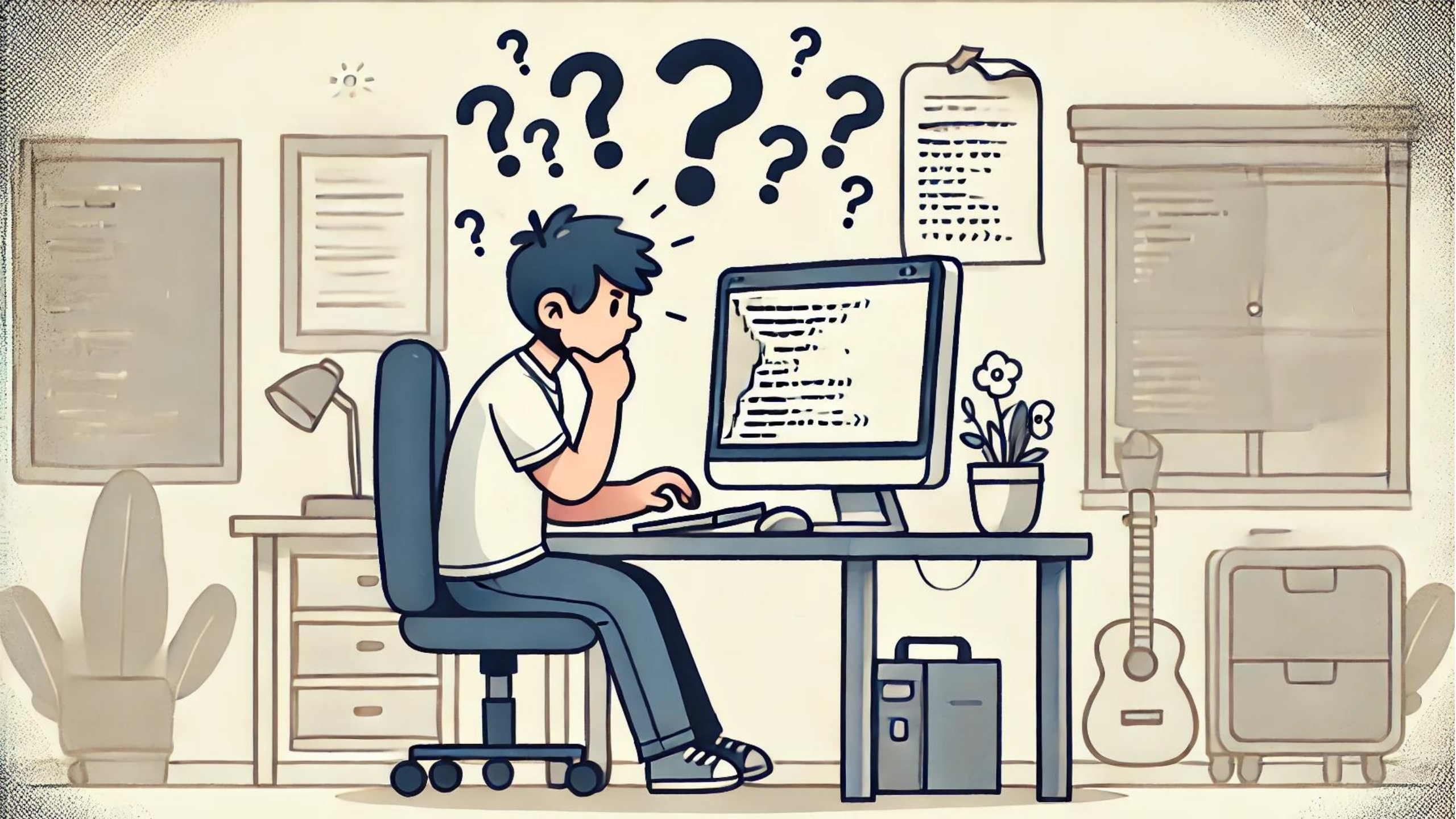
The route optimization workflow demonstrates how to use NVIDIA cuOpt to minimize vehicle routing inefficiencies by finding the most optimal route for a fleet of vehicles making deliveries, pickups, dispatching jobs, etc.

[Browse >](#)

NVIDIA DOCS HUB



[Link](#)





Christopher Kanagaraj

@Chrissuccess



Tom Cruise's stunt doubles 🤖



8:32 PM · Jun 7, 2023 · 1M Views