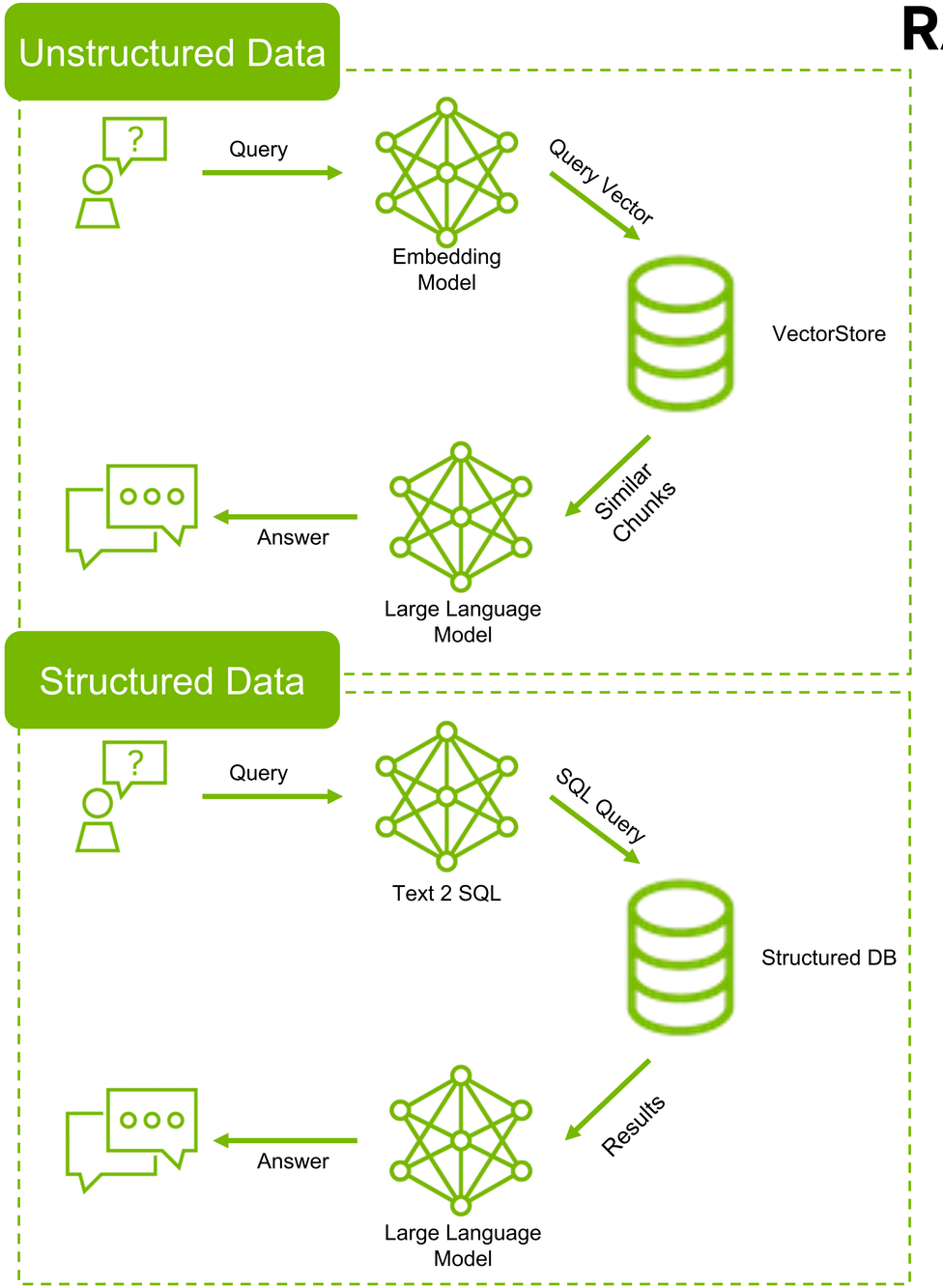
Beyond RAG: Building Agents, CoPilots, Assistants and More!

Tanay Varshney, Senior Developer Advocate Engineer - Deep Learning
Annie Surla, Developer Advocate Engineer - Deep Learning
Aditi Bodhankar, Developer Advocate Engineer - Deep Learning

GTC’24 - March 18th, 2024

RAG - What can it do?



NVIDIA Announces Financial Results for Third Quarter Fiscal 2024

What are the GAAP earnings per Diluted share for Q3 2024?

November 21, 2023

- Record revenue of $18.12 billion, up 34% from Q2, up 206% from year ago
- Record Data Center revenue of $14.51 billion, up 41% from Q2, up 279% from year ago

NVIDIA (NASDAQ: NVDA) today reported revenue for the third quarter ended October 29, 2023, of $18.12 billion, up 206% from a year ago and up 34% from the previous quarter.

GAAP earnings per diluted share for the quarter were $3.71, up more than 12x from a year ago and up 50% from the previous quarter. Non-GAAP earnings per diluted share were $4.02, up nearly 6x from a year ago and up 49% from the previous quarter.

"Our strong growth reflects the broad industry platform transition from general-purpose to accelerated computing and generative AI," said Jensen Huang, founder and CEO of NVIDIA.

"Large language model startups, consumer internet companies and global cloud service providers were the first movers, and the next waves are starting to build. Nations and regional CSPs are investing in AI clouds to serve local demand, enterprise software companies are adding AI copilots and assistants to their platforms, and enterprises are creating custom AI to automate the world's largest industries.

"NVIDIA GPUs, CPUs, networking, AI foundry services and NVIDIA AI Enterprise software are all growth engines in full throttle. The era of generative AI is taking off," he said.

NVIDIA will pay its next quarterly cash dividend of $0.04 per share on December 28, 2023, to all shareholders of record on December 6, 2023.

Q3 Fiscal 2024 Summary

	GAAP				
($ in millions, except earnings per share)	Q3 FY24	Q2 FY24	Q3 FY23	Q/Q	Y/Y
Revenue	$18,120	$13,507	$5,931	Up 34%	Up 206%
Gross margin	74.0%	70.1%	53.6%	Up 3.9 pts	Up 20.4 pts
Operating expenses	$2,983	$2,662	$2,576	Up 12%	Up 16%
Operating income	$10,417	$6,800	$601	Up 53%	Up 1,633%
Net income	$9,243	$6,188	$680	Up 49%	Up 1,259%
Diluted earnings per share	$3.71	$2.48	$0.27	Up 50%	Up 1,274%

RAG - Is it enough?

NVIDIA Announces Financial Results for Third Quarter Fiscal 2024

November 21, 2023

- Record revenue of $18.12 billion, up 34% from Q2, up 206% from year ago
- Record Data Center revenue of $14.51 billion, up 41% from Q2, up 279% from year ago

NVIDIA (NASDAQ: NVDA) today reported revenue for the third quarter ended October 29, 2023, of $18.12 billion, up 206% from a year ago and up 34% from the previous quarter.

GAAP earnings per diluted share for the quarter were $3.71, up more than 12x from a year ago and up 50% from the previous quarter. Non-GAAP earnings per diluted share were $4.02, up nearly 6x from a year ago and up 49% from the previous quarter.

"Our strong growth reflects the broad industry platform transition from general-purpose to accelerated computing and generative AI," said Jensen Huang, founder and CEO of NVIDIA.

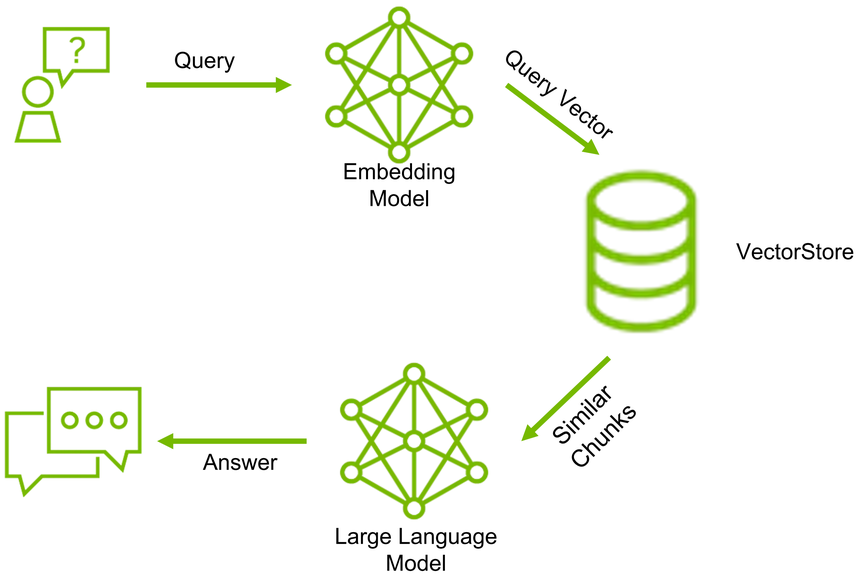
"Large language model startups, consumer internet companies and global cloud service providers were the first movers, and the next waves are starting to build. Nations and regional CSPs are investing in AI clouds to serve local demand, enterprise software companies are adding AI copilots and assistants to their platforms, and enterprises are creating custom AI to automate the world's largest industries.

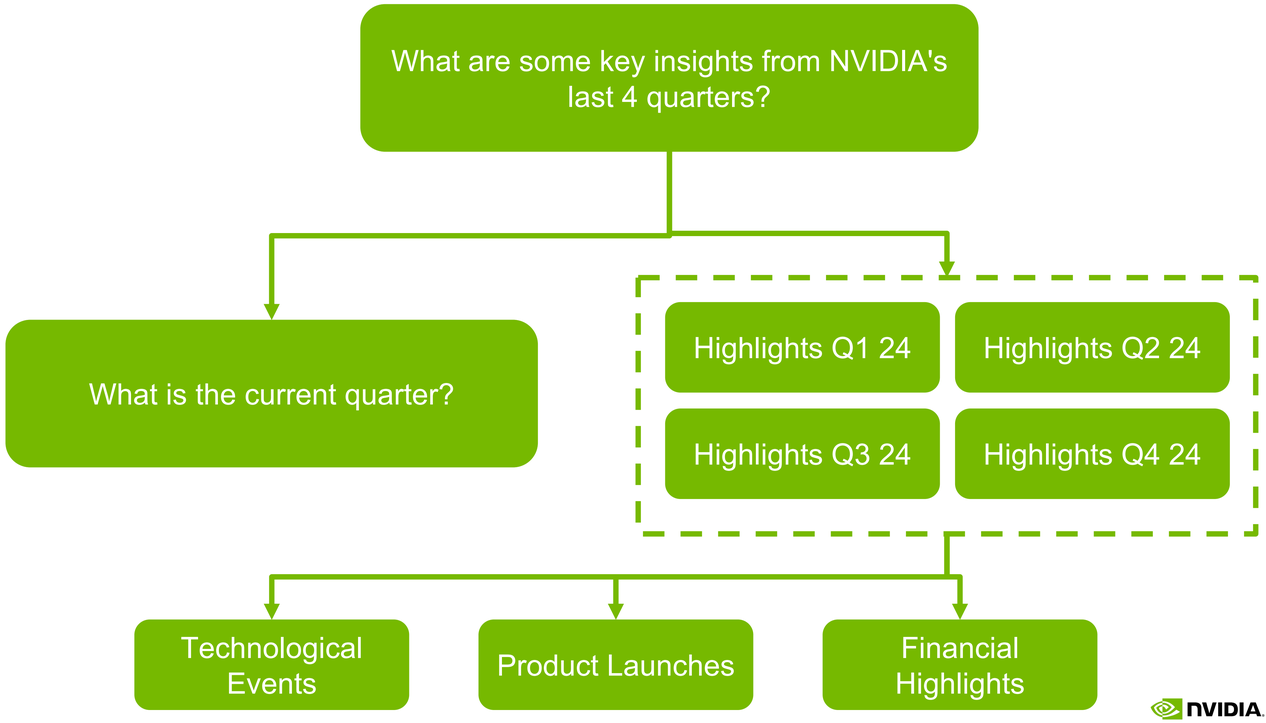
"NVIDIA GPUs, CPUs, networking, AI foundry services and NVIDIA AI Enterprise software are all growth engines in full throttle. The era of generative AI is taking off," he said.

NVIDIA will pay its next quarterly cash dividend of $0.04 per share on December 28, 2023, to all shareholders of record on December 6, 2023.

Q3 Fiscal 2024 Summary

	GAAP				
($ in millions, except earnings per share)	Q3 FY24	Q2 FY24	Q3 FY23	Q/Q	Y/Y
Revenue	$18,120	$13,507	$5,931	Up 34%	Up 206%
Gross margin	74.0%	70.1%	53.6%	Up 3.9 pts	Up 20.4 pts
Operating expenses	$2,983	$2,662	$2,576	Up 12%	Up 16%
Operating income	$10,417	$6,800	$601	Up 53%	Up 1,633%
Net income	$9,243	$6,188	$680	Up 49%	Up 1,259%
Diluted earnings per share	$3.71	$2.48	$0.27	Up 50%	Up 1,274%





Is “Search” all you need?

NVIDIA Announces Financial Results for Third Quarter Fiscal 2024

November 21, 2023

- Record revenue of $18.12 billion, up 34% from Q2, up 206% from year ago
- Record Data Center revenue of $14.51 billion, up 41% from Q2, up 279% from year ago

NVIDIA (NASDAQ: NVDA) today reported revenue for the third quarter ended October 29, 2023, of $18.12 billion, up 206% from a year ago and up 34% from the previous quarter.

GAAP earnings per diluted share for the quarter were $3.71, up more than 12x from a year ago and up 50% from the previous quarter. Non-GAAP earnings per diluted share were $4.02, up nearly 6x from a year ago and up 49% from the previous quarter.

“Our strong growth reflects the broad industry platform transition from general-purpose to accelerated computing and generative AI,” said Jensen Huang, founder and CEO of NVIDIA.

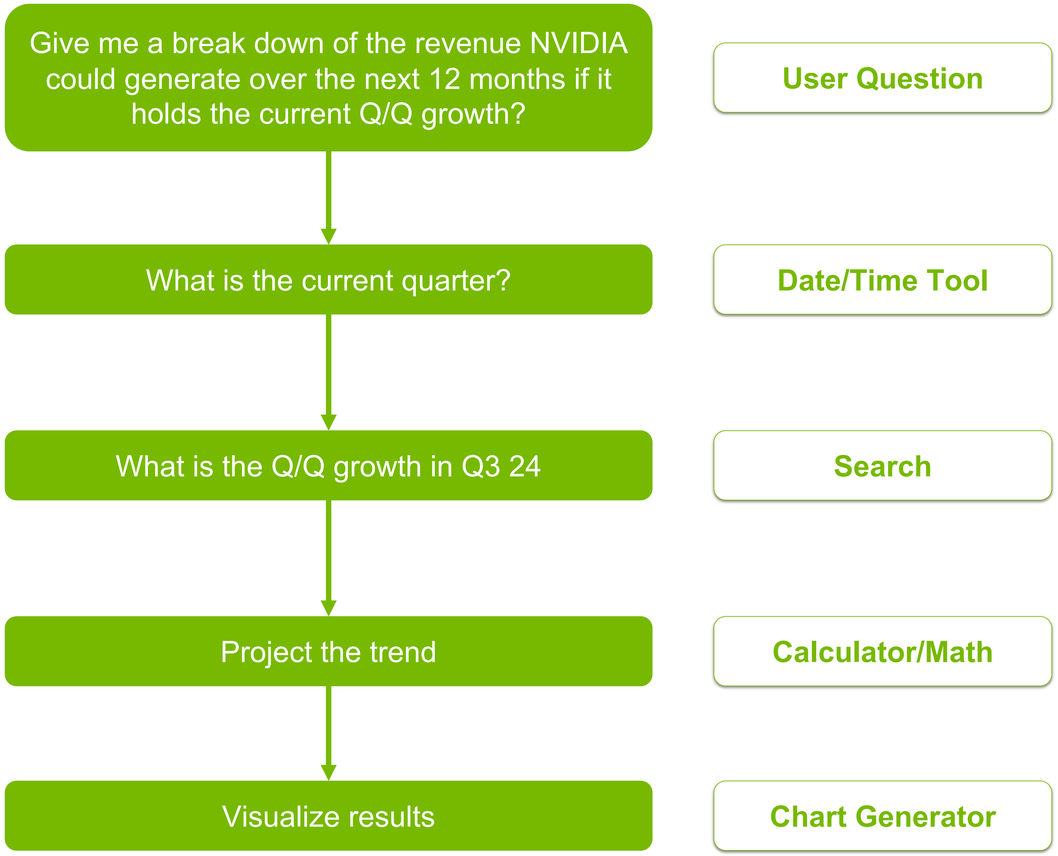
“Large language model startups, consumer internet companies and global cloud service providers were the first movers, and the next waves are starting to build. Nations and regional CSPs are investing in AI clouds to serve local demand, enterprise software companies are adding AI copilots and assistants to their platforms, and enterprises are creating custom AI to automate the world’s largest industries.

“NVIDIA GPUs, CPUs, networking, AI foundry services and NVIDIA AI Enterprise software are all growth engines in full throttle. The era of generative AI is taking off,” he said.

NVIDIA will pay its next quarterly cash dividend of $0.04 per share on December 28, 2023, to all shareholders of record on December 6, 2023.

Q3 Fiscal 2024 Summary

	GAAP				
($ in millions, except earnings per share)	Q3 FY24	Q2 FY24	Q3 FY23	Q/Q	Y/Y
Revenue	$18,120	$13,507	$5,931	Up 34%	Up 206%
Gross margin	74.0%	70.1%	53.6%	Up 3.9 pts	Up 20.4 pts
Operating expenses	$2,983	$2,662	$2,576	Up 12%	Up 16%
Operating income	$10,417	$6,800	$601	Up 53%	Up 1,633%
Net income	$9,243	$6,188	$680	Up 49%	Up 1,259%
Diluted earnings per share	$3.71	$2.48	$0.27	Up 50%	Up 1,274%



Is “Search” all you need? (contd)

I need to buy a toothpaste, a couple of chocolate bars, and AAA batteries.

User Question

Locate “Toothpaste”, “chocolate bar”, “AAA Battery”

Search

Option 1: Provide the “Shortest path”

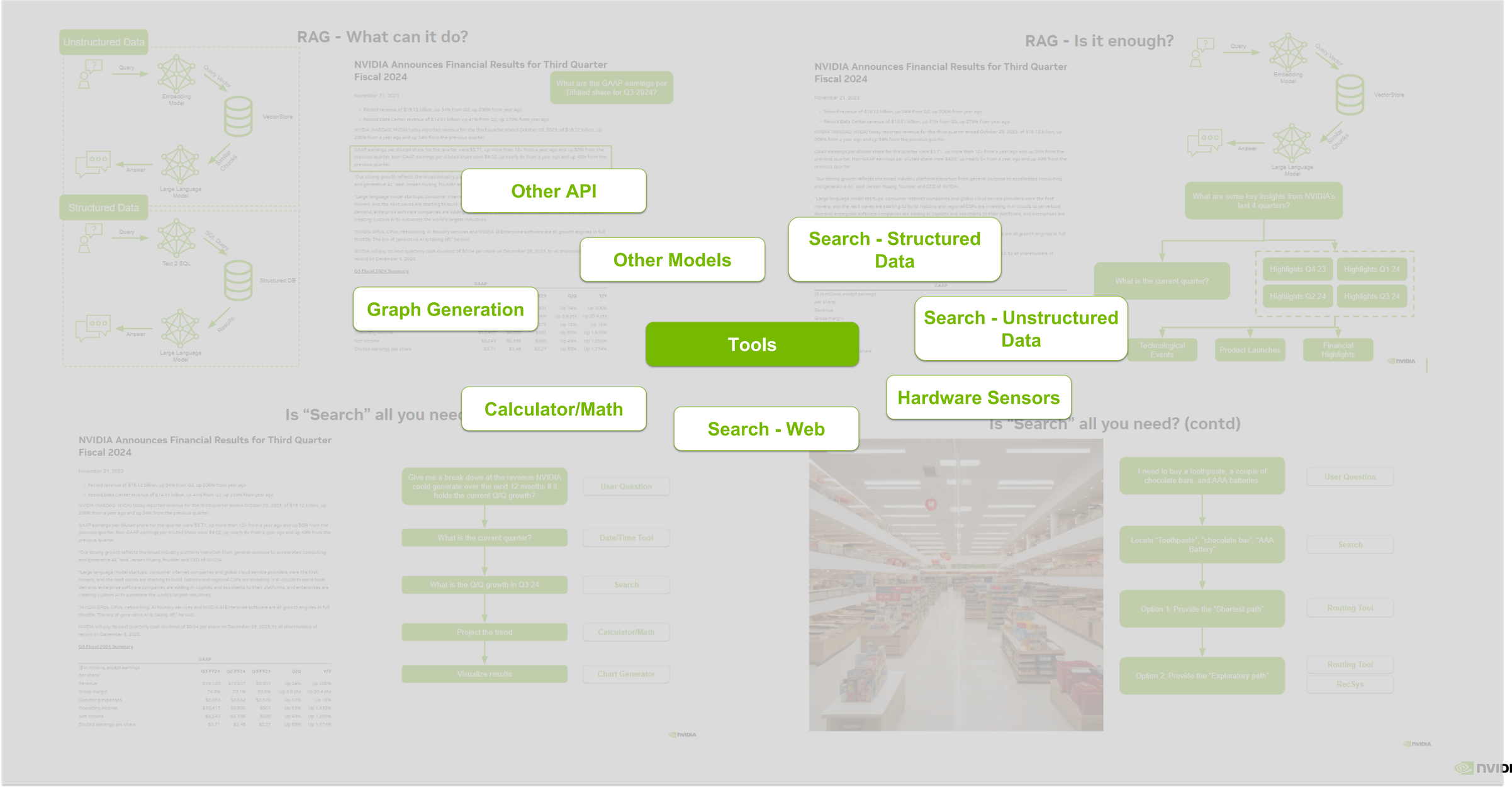
Routing Tool

Option 2: Provide the “Exploratory path”

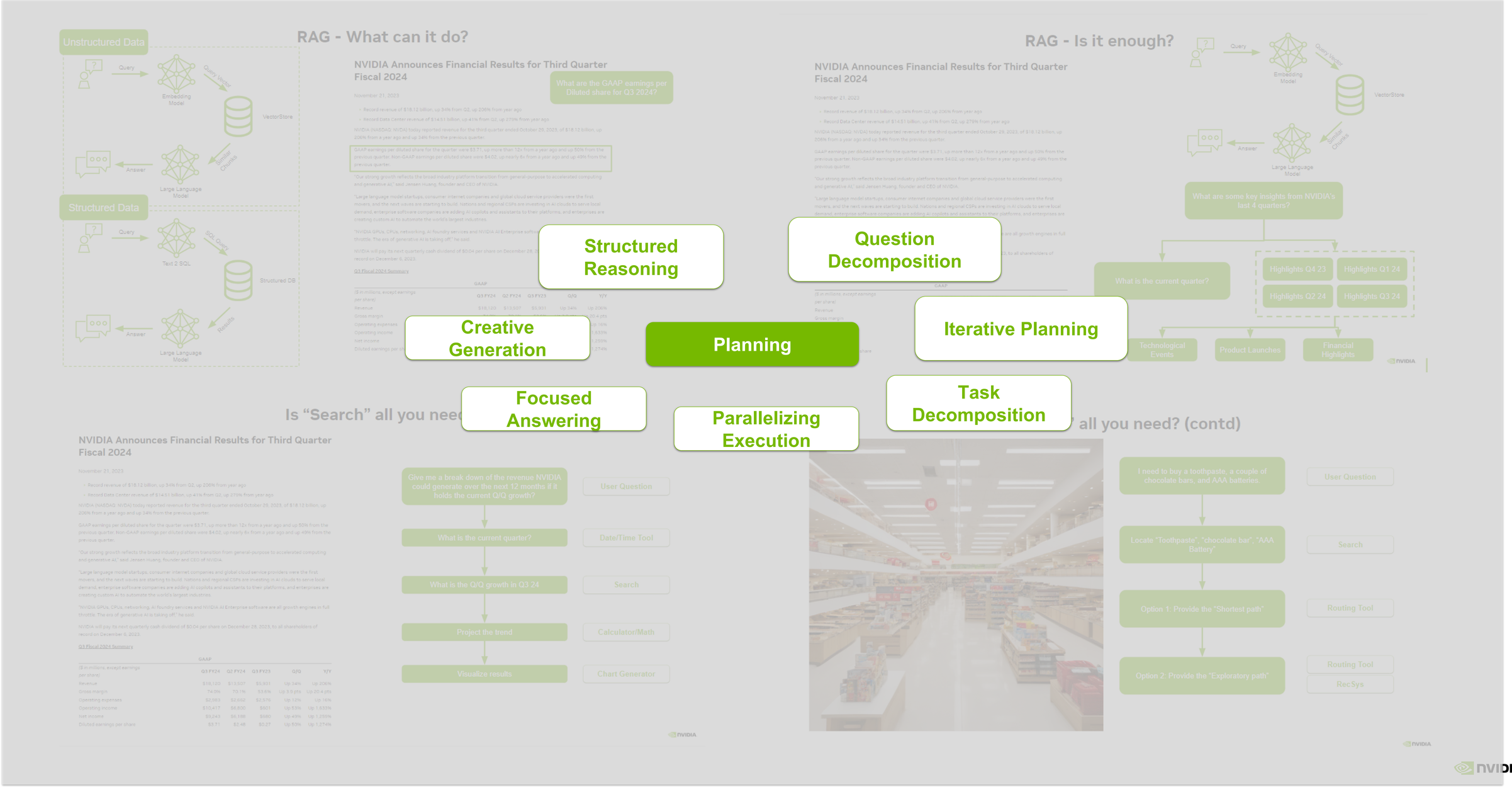
Routing Tool

RecSys

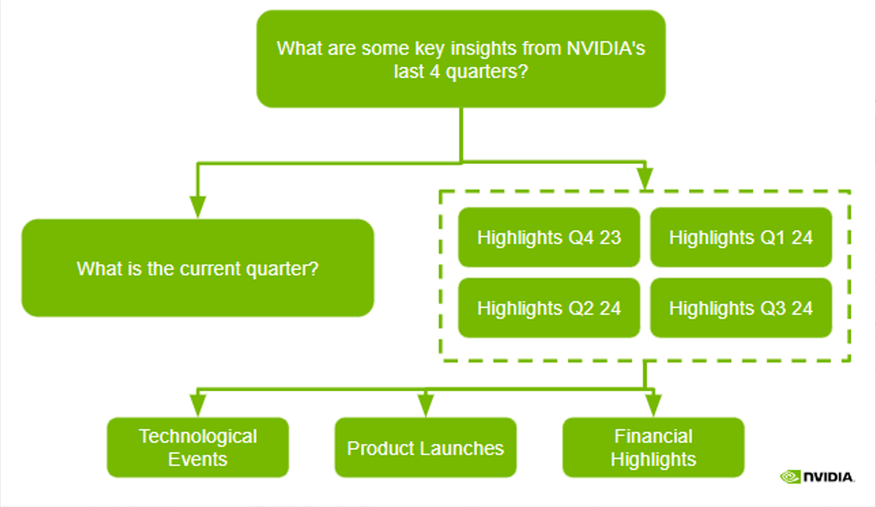
What do we need to solve complex questions?



What do we need to solve complex questions?



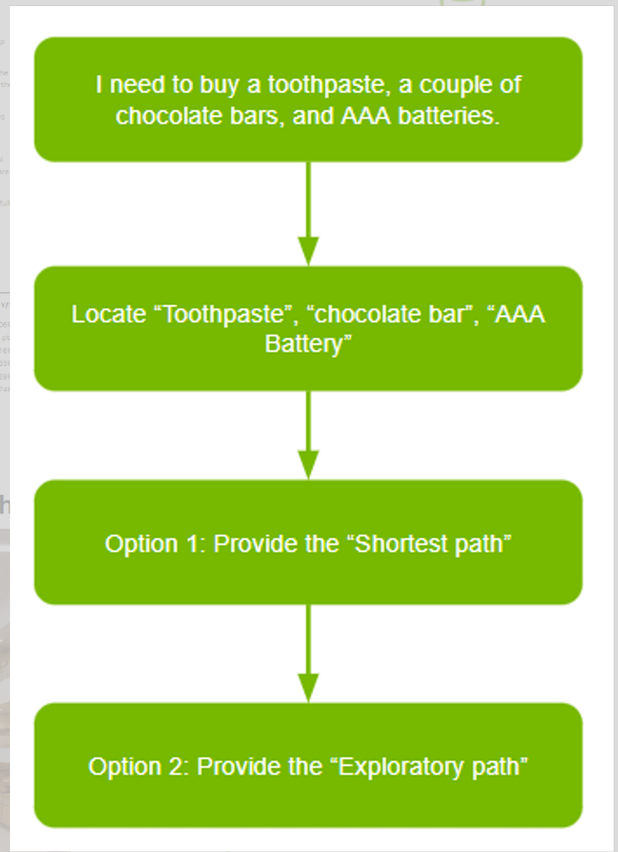
What do we need to solve complex questions?



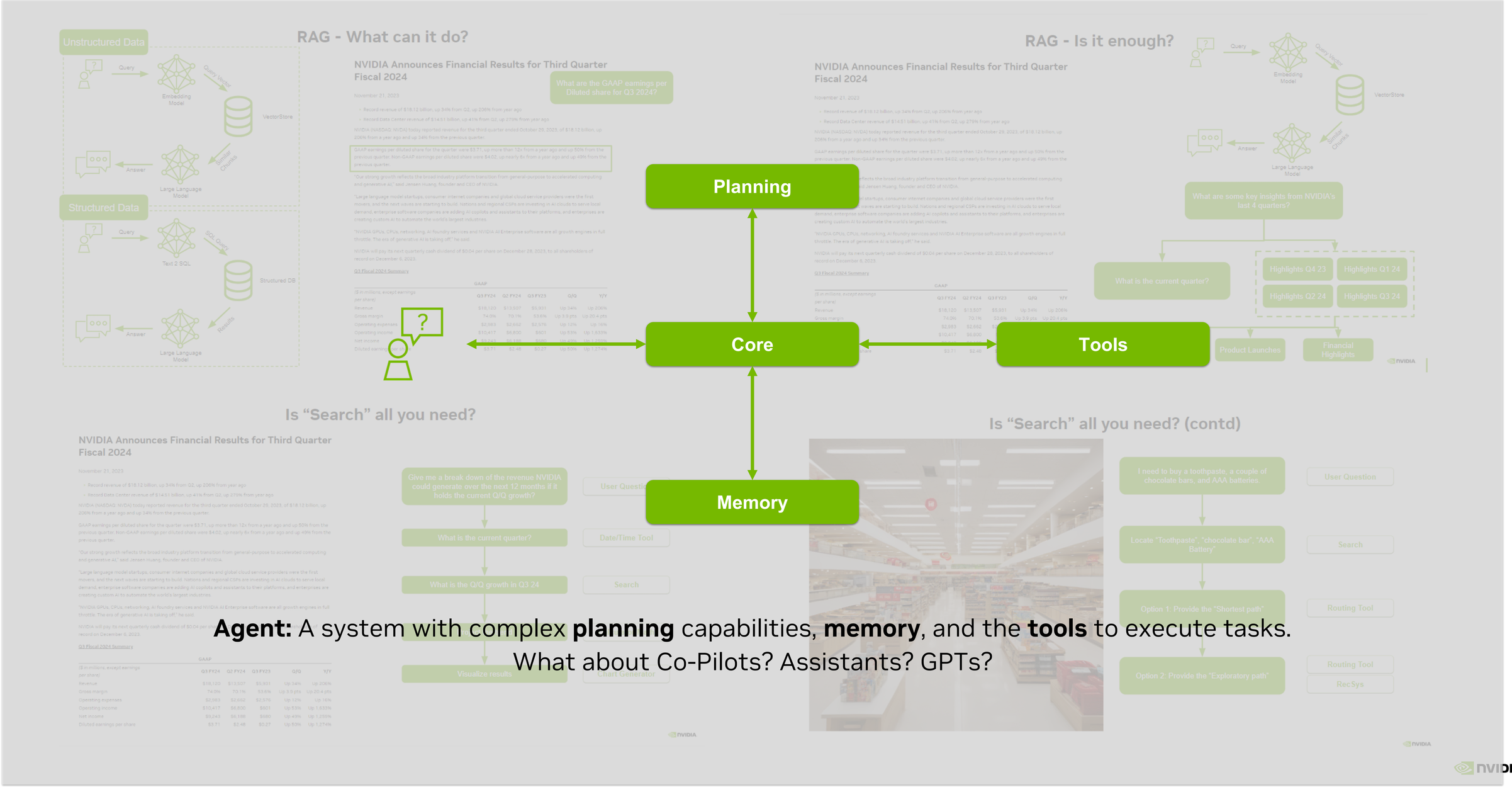
Memory

Long Term Memory

Short Term Memory



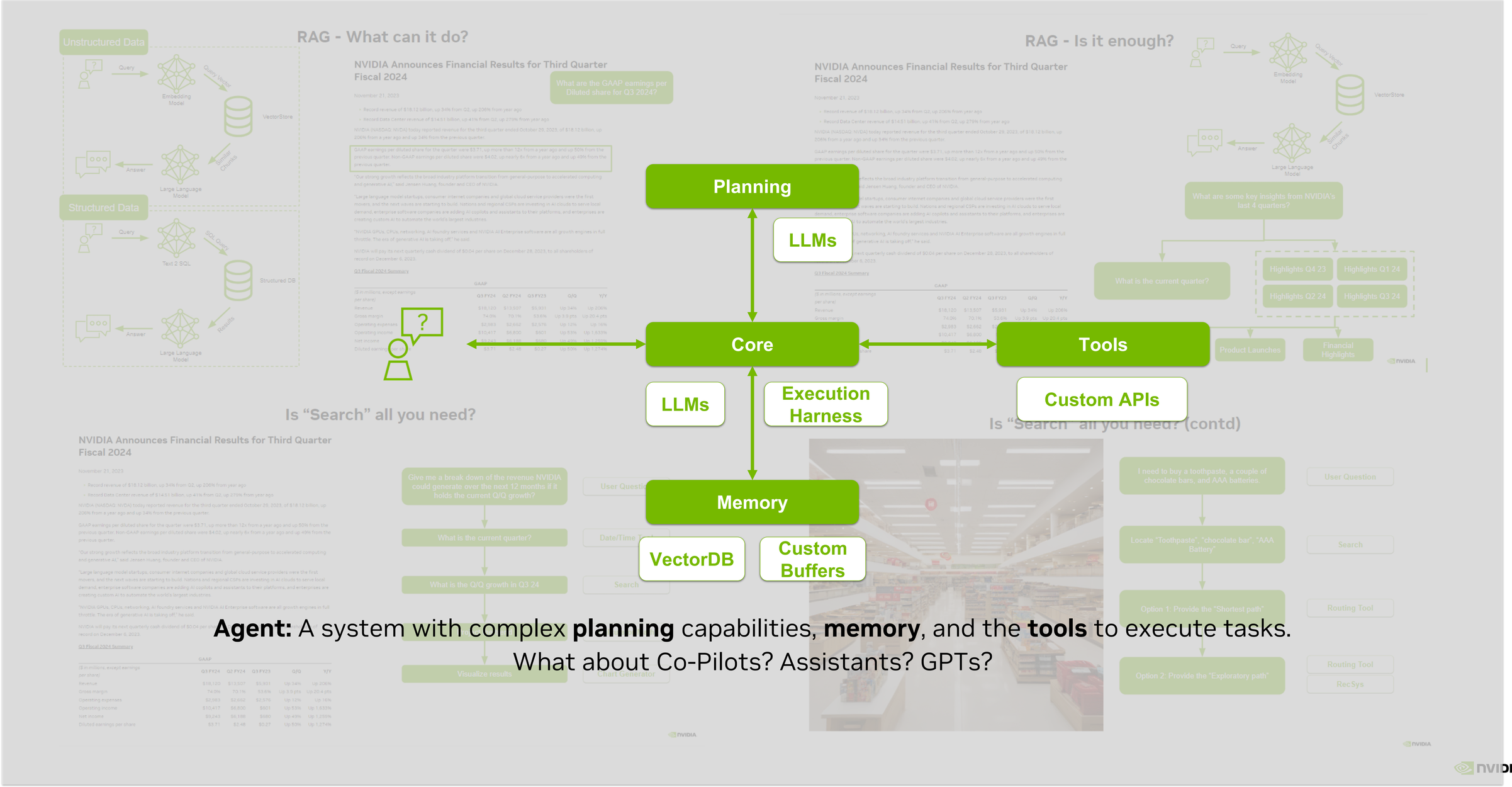
Agent - A general structure



Agent: A system with complex **planning** capabilities, **memory**, and the **tools** to execute tasks.

What about Co-Pilots? Assistants? GPTs?

Agent - A general structure



Agent: A system with complex **planning** capabilities, **memory**, and the **tools** to execute tasks.

What about Co-Pilots? Assistants? GPTs?

Building your first agent

NVIDIA Corporation (NASDAQ:NVDA) Q3 2024 Earnings Conference Call
November 21, 2023 5:00 PM ET

Company Participants

Simona Jankowski - VP, IR

Colette Kress - EVP & CFO

Jensen Huang - President & CEO

Conference Call Participants

Vivek Arya - Bank of America

Aaron Rakers - Wells Fargo

Joseph Moore - Morgan Stanley

Tim Arcuri - UBS

Toshiya Hari - Goldman Sachs

Stacy Rasgon - Bernstein Research

Matt Ramsay - Cowen

Harlan Sur - J.P. Morgan

Operator

Good afternoon. My name is JL, and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's Third Quarter Earnings Call. All lines have been placed on mute to prevent any background noise. After the speakers' remarks, there will be a question-and-answer session. [Operator Instructions]

Simona Jankowski, you may now begin your conference.

Simona Jankowski

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the third quarter of fiscal 2024. With me today from NVIDIA are Jensen Huang, President and Chief Executive Officer; and Colette Kress, Executive Vice President and Chief Financial Officer. I'd like to remind you that our call is being webcast live on NVIDIA's Investor Relations website. The webcast will be available for replay until the conference call to discuss our financial results for the fourth quarter and fiscal 2024. The content of today's call is NVIDIA's property. It can't be reproduced or transcribed without our prior written consent.

During this call, we may make forward-looking statements based on current expectations. These are subject to a number of significant risks and uncertainties and our actual results may differ materially. For a discussion of factors that could affect our future financial results and business, please

During this call, we may make forward-looking statements based on current expectations. These are subject to a number of significant risks and uncertainties and our actual results may differ materially. For a discussion of factors that could affect our future financial results and business, please refer to the disclosure in today's earnings release, our most recent forms 10-K and 10-Q, and the reports that we may file on Form 8-K with the Securities and Exchange Commission. All statements are made as of today, November 21, 2023, based on information currently available to us. Except as required by law, we assume no obligation to update any such statements.

During this call, we will discuss non-GAAP financial measures. You can find a reconciliation of these non-GAAP financial measures to GAAP financial measures in our CFO commentary, which is posted on our website.

With that, let me turn the call over to Colette.

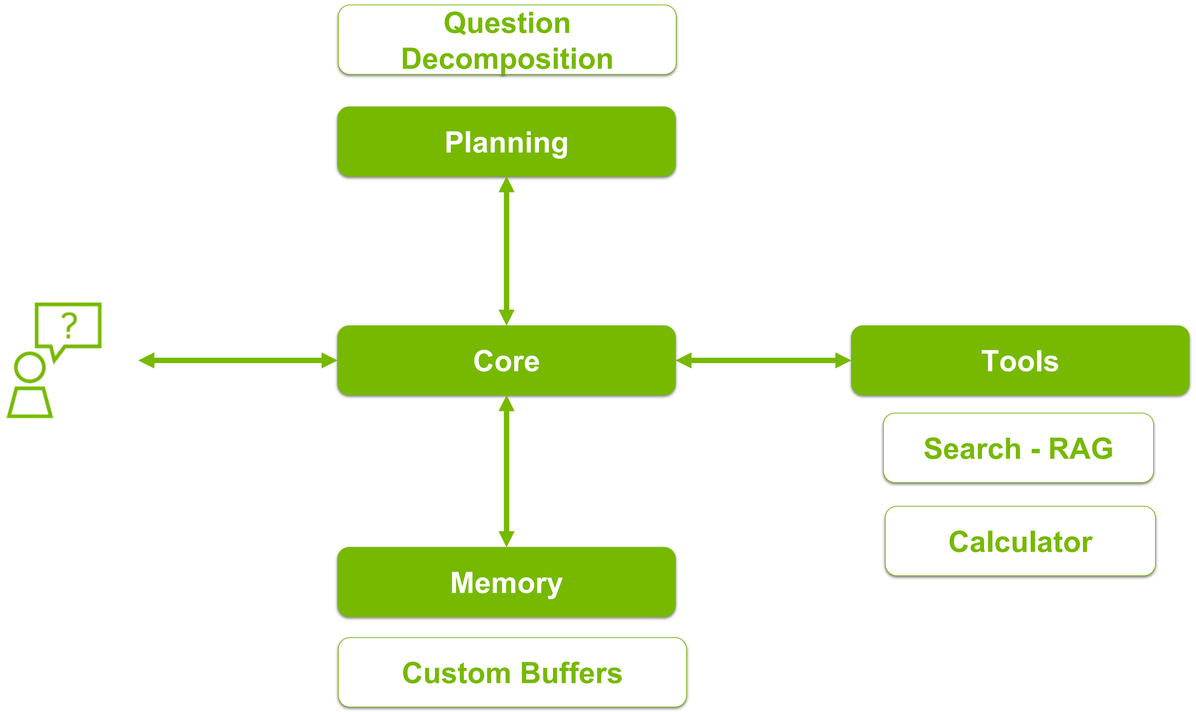
Colette Kress

Thanks, Simona. Q3 was another record quarter. Revenue of $18.1 billion was up 34% sequentially and up more than 200% year-on-year and well above our outlook for $16 billion. Starting with Data Center. The continued ramp of the NVIDIA HGX platform based on our Hopper Tensor Core GPU architecture, along with InfiniBand end-to-end networking, drove record revenue of $14.5 billion, up 41% sequentially and up 279% year-on-year.

NVIDIA HGX with InfiniBand together are essentially the reference architecture for AI supercomputers and data center infrastructures. Some of the most exciting generative AI applications are built and run on NVIDIA, including Adobe Firefly, ChatGPT, Microsoft 365 Copilot, CoAssist, now assist with ServiceNow and Zoom AI Companion. Our Data Center compute revenue quadrupled from last year and networking revenue nearly tripled.

Investments in infrastructure for training and inferencing large language models, deep learning, recommender systems and generative AI applications is fueling strong broad-based demand for NVIDIA accelerated computing. Inferencing is now a major workload for NVIDIA AI computing. Consumer Internet companies and enterprises drove exceptional sequential growth in Q3, comprising approximately half of our Data Center revenue and outpacing total growth.

Companies like Meta are in full production with deep learning, recommender systems and also investing in generative AI to help advertisers optimize images and text. Most major consumer Internet companies are racing to ramp up generative AI deployment. The enterprise wave of AI adoption is now beginning. Enterprise software companies such as Adobe, Databricks, Snowflake and ServiceNow are adding AI copilots and the systems to their platforms. And broader enterprises are developing custom AI for vertical industry applications such as Tesla in autonomous



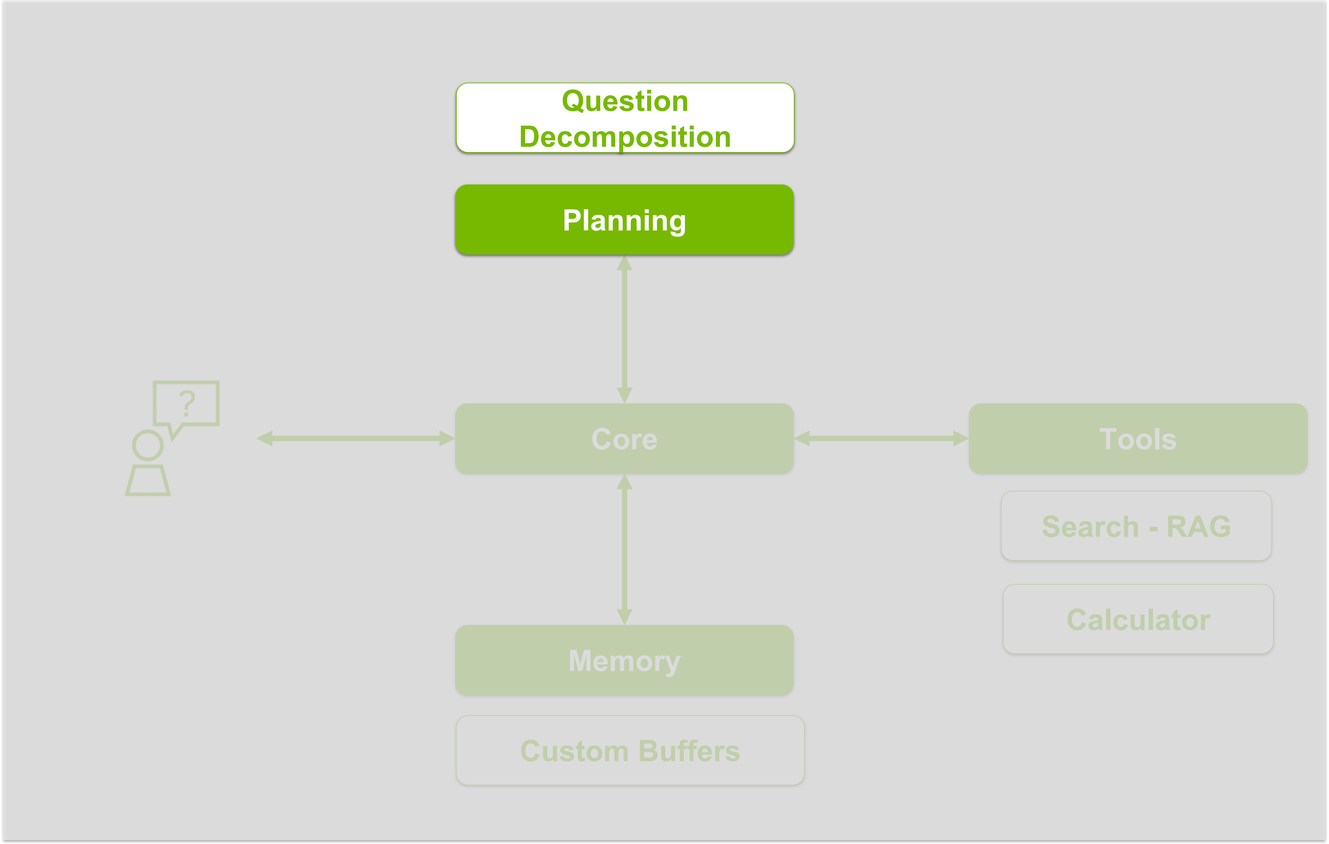
Question: How much did revenue increase by between Q3 of FY 2024 and Q1 of FY 2024?

Building your first agent - Planning

```
"""GENERAL INSTRUCTIONS
You are a domain expert. Your task is to
break down a complex question into simpler
sub-parts.

USER QUESTION
How much did revenue increase by between Q3
of FY 2024 and Q1 of FY 2024?

ANSWER FORMAT
{"sub-questions":["<FILL>"]}"""
```



M {"sub-questions":["What is the revenue for Q3 of FY 2024?", "What is the revenue for Q1 of FY 2024?"]}

Question: How much did revenue increase by between Q3 of FY 2024 and Q1 of FY 2024?

Building your first agent - Tools

NVIDIA Generative AI Examples

Search - RAG

Introduction

State-of-the-art Generative AI examples that are easy to deploy, test, and extend. All examples run on the high performance NVIDIA CUDA-X software stack and NVIDIA GPUs.

NVIDIA NGC

Generative AI Examples uses resources from the NVIDIA NGC AI Development Catalog.

Sign up for a free NGC developer account to access:

- GPU-optimized containers used in these examples
- Release notes and developer documentation

Retrieval Augmented Generation (RAG)

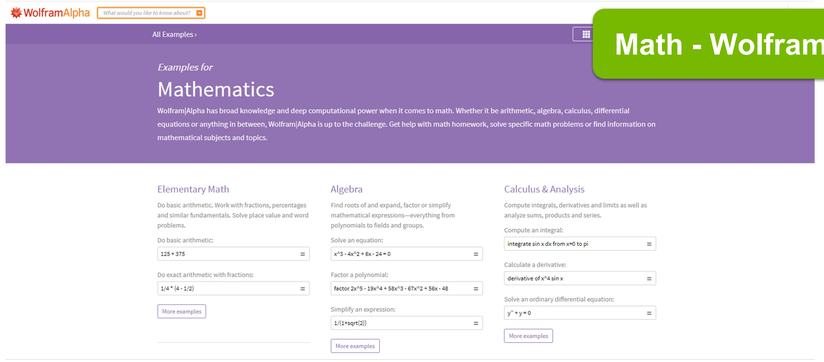
A RAG pipeline embeds multimodal data -- such as documents, images, and video -- into a database connected to a LLM. RAG lets users chat with their data!

Developer RAG Examples

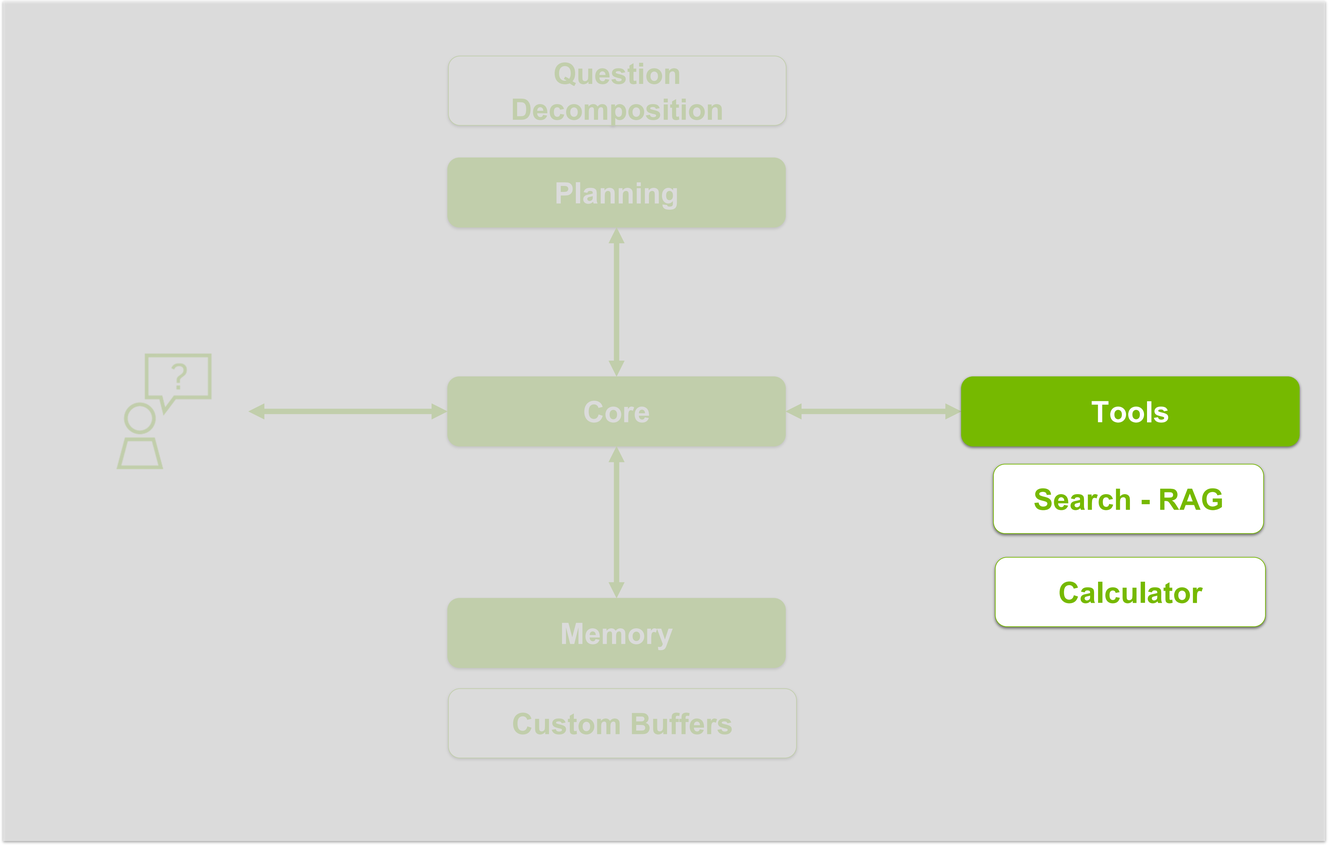
The developer RAG examples run on a single VM. They demonstrate how to combine NVIDIA GPU acceleration with popular LLM programming frameworks using NVIDIA's open source connectors. The examples are easy to deploy via Docker Compose.

Examples support local and remote inference endpoints. If you have a GPU, you can inference locally via TensorRT-LLM. If you don't have a GPU, you can inference and embed remotely via NVIDIA AI Foundations endpoints.

Model	Embedding	Framework	Description	Multi-GPU	TRT-LLM	NVIDIA AI Foundation	Triton	
llama-2	e5-large-v2	Llamaindex	Canonical QA Chatbot	YES	YES	No	YES	Mil



Math - WolframAlpha



Question: How much did revenue increase by between Q3 of FY 2024 and Q1 of FY 2024?

Building your first agent - Core

```
GENERAL INSTRUCTIONS: Your task is to answer
questions. If you cannot answer the question,
request a helper or use a tool. Fill with Nil where
no tool or helper is required.

<TOOLS>
- Search Tool
- Math Tool
</TOOLS>

<HELPERS>
- Decomposition: Breaks Complex Questions down into
simpler subparts
</HELPERS>

<CONTEXTUAL_INFORMATION>
</CONTEXTUAL_INFORMATION>

QUESTION: How much did the revenue grow between Q1
of 2024 and Q2 of 2024?

ANSWER FORMAT: {"Tool_Request": "<Fill>",
"Helper_Request "<Fill>"}
```

```
def Agent(Question, Context):
    Action = LLM(Context + Question)

    if Action == "Decomposition":
        Sub Questions = LLM(Question)
        Agent(Sub Question, Context)

    if Action == "Search Tool":
        Answer = RAG_Pipeline(Question)
        Context = Context + Answer
        Agent(Question, Context)

    if Action == "Gen Final Answer”:
        return LLM(Context)

    if Action == "<Another Tool>":
        <Execute Another Tool>
```

Building your first agent - Workflow

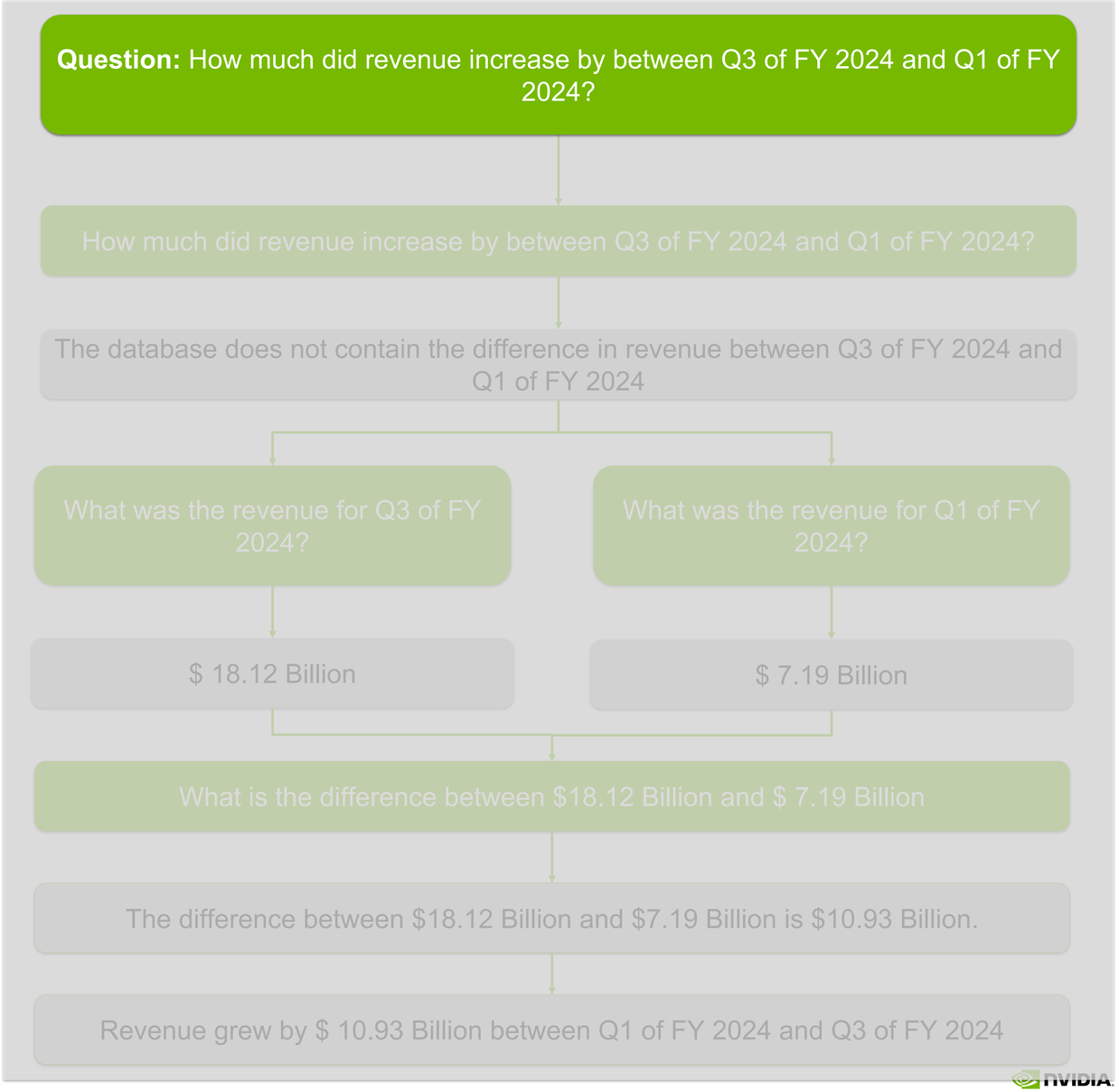
```
def Agent(Question, Context):
    Action = LLM(Context + Question)

    if Action == "Decomposition":
        Sub Questions = LLM(Question)
        Agent(Sub Question, Context)

    if Action == "Search Tool":
        Answer = RAG_Pipeline(Question)
        Context = Context + Answer
        Agent(Question, Context)

    if Action == "Gen Final Answer”:
        return LLM(Context)

    if Action == "Math Tool":
        math_tool(Question)
```



Building your first agent - Workflow

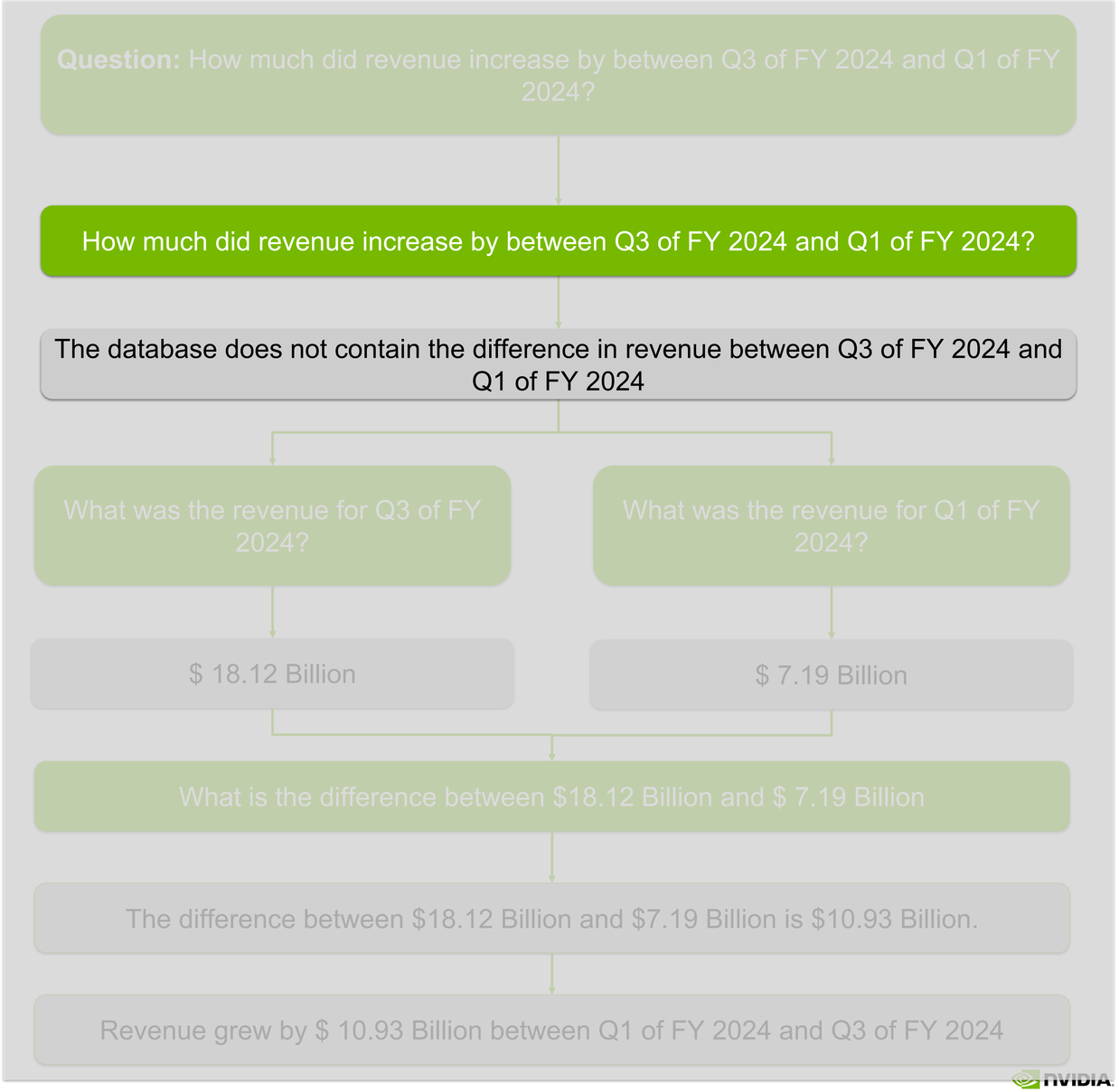
```
def Agent(Question, Context):
    Action = LLM(Context + Question)

    if Action == "Decomposition":
        Sub Questions = LLM(Question)
        Agent(Sub Question, Context)

    if Action == "Search Tool":
        Answer = RAG_Pipeline(Question)
        Context = Context + Answer
        Agent(Question, Context)

    if Action == "Gen Final Answer”:
        return LLM(Context)

    if Action == "Math Tool":
        math_tool(Question)
```



Building your first agent - Workflow

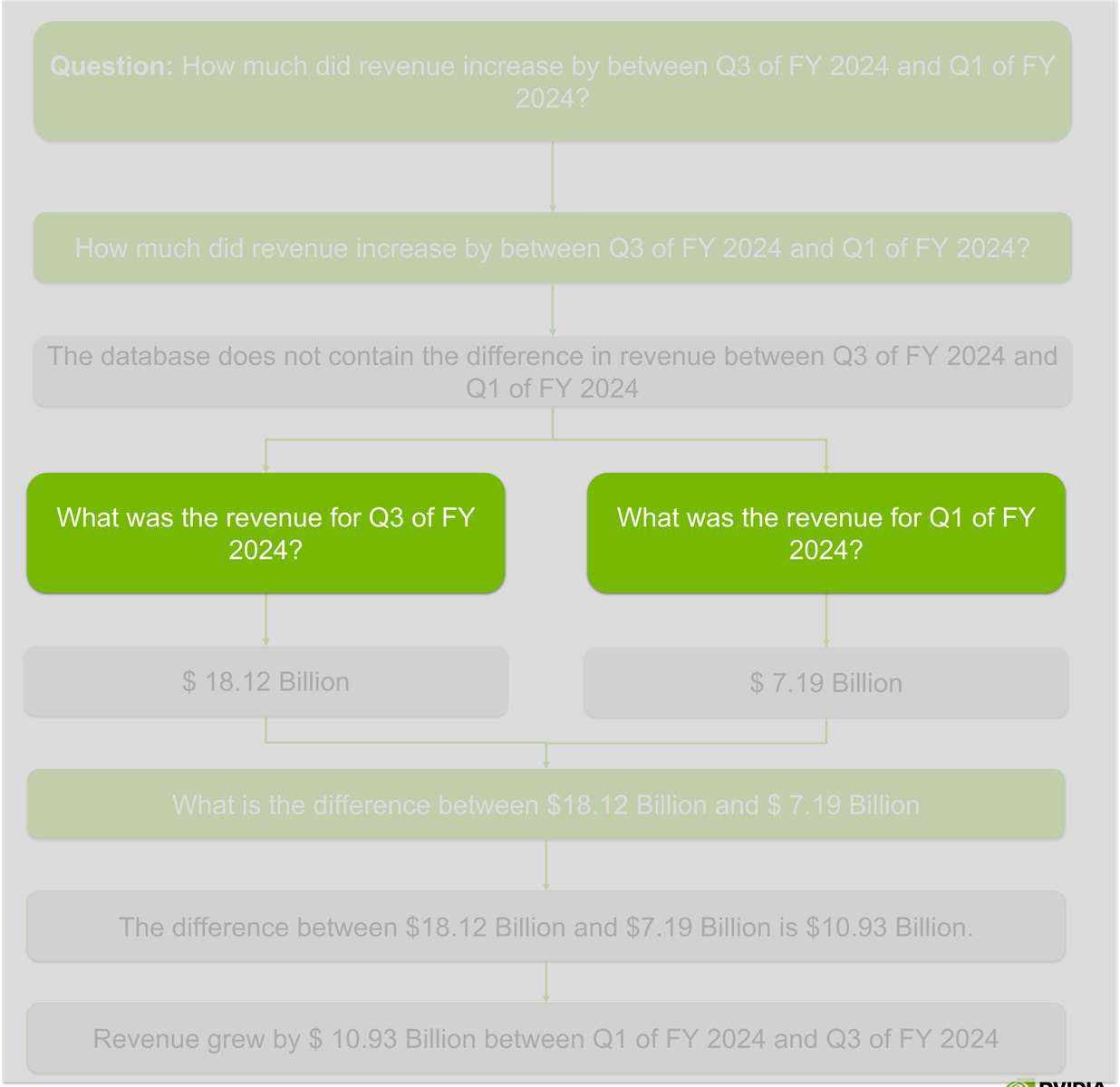
```
def Agent(Question, Context):
    Action = LLM(Context + Question)

    if Action == "Decomposition":
        Sub Questions = LLM(Question)
        Agent(Sub Question, Context)

    if Action == "Search Tool":
        Answer = RAG_Pipeline(Question)
        Context = Context + Answer
        Agent(Question, Context)

    if Action == "Gen Final Answer”:
        return LLM(Context)

    if Action == "Math Tool":
        math_tool(Question)
```



Building your first agent - Workflow

```
def Agent(Question, Context):
    Action = LLM(Context + Question)

    if Action == "Decomposition":
        Sub Questions = LLM(Question)
        Agent(Sub Question, Context)

    if Action == "Search Tool":
        Answer = RAG_Pipeline(Question)
        Context = Context + Answer
        Agent(Question, Context)

    if Action == "Gen Final Answer”:
        return LLM(Context)

    if Action == "Math Tool":
        math_tool(Question)
```

Building your first agent - Workflow

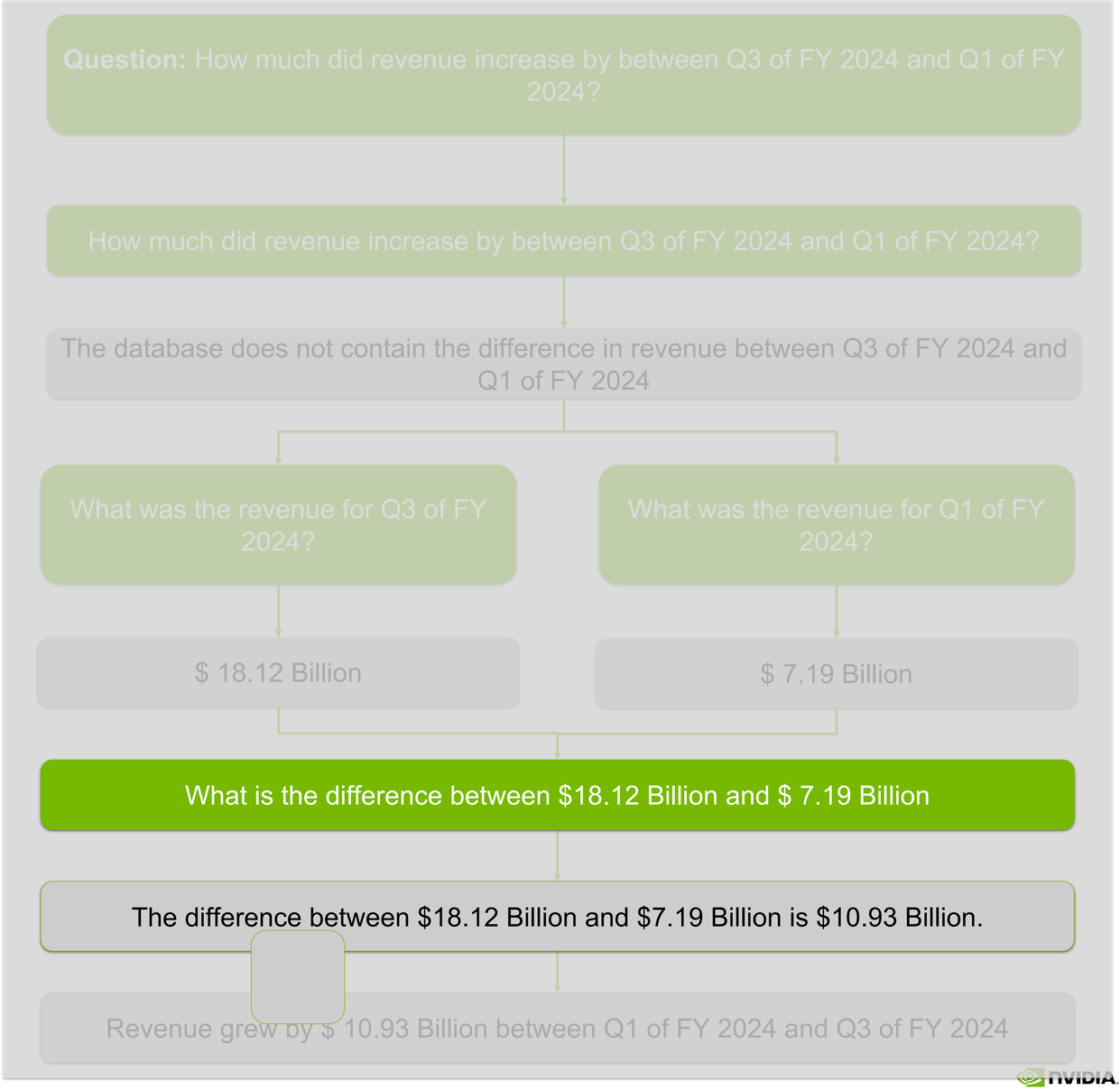
```
def Agent(Question, Context):
    Action = LLM(Context + Question)

    if Action == "Decomposition":
        Sub Questions = LLM(Question)
        Agent(Sub Question, Context)

    if Action == "Search Tool":
        Answer = RAG_Pipeline(Question)
        Context = Context + Answer
        Agent(Question, Context)

    if Action == "Gen Final Answer”:
        return LLM(Context)

    if Action == "Math Tool":
        math_tool(Question)
```



Building your first agent - Workflow

```
def Agent(Question, Context):
    Action = LLM(Context + Question)

    if Action == "Decomposition":
        Sub Questions = LLM(Question)
        Agent(Sub Question, Context)

    if Action == "Search Tool":
        Answer = RAG_Pipeline(Question)
        Context = Context + Answer
        Agent(Question, Context)

    if Action == "Gen Final Answer”:
        return LLM(Context)

    if Action == "Math Tool":
        math_tool(Question)
```

Spectrum of Agents

Extractive — Spectrum of Agents — Executive

Data Agents

API Agents

How many quarters did this company have a positive cash flow this year?

- Ask questions or extract information from a plethora of data sources
- Ask for assistive reasoning tasks

Write a tweet for ad campaign around selling more detergent. Also **generate an image** to go with it. While you are at it also **add** the new product has a softner in the tweet. **Brainstorm some other ideas** for marketing..

- Execute what a user (human) or another agent (machine) requires
- Traditionally done via APIs, application logic and interaction layer like a “Web Application/Page”

Data Agent

Use Case: Talking to a Relational DB for Inventory Management

Schema of a simple database to showcase the use case

Inventory Table

Inventory ID	Product ID	Quantity in Stock	Min Stock
1	1	150	30

Product/SKU Table

Product ID	Name	Description	Price	Supplier ID
1	iPhone 15 Pro	Apple Flagship Smartphone	999.00	2

Supplier Table

Supplier ID	Name	Address	Contact
2	Apple Inc.	Cupertino, California, USA	800-692-7753

Disclaimer: - The information above are just placeholders for an example and not intended to reflect the actual details.

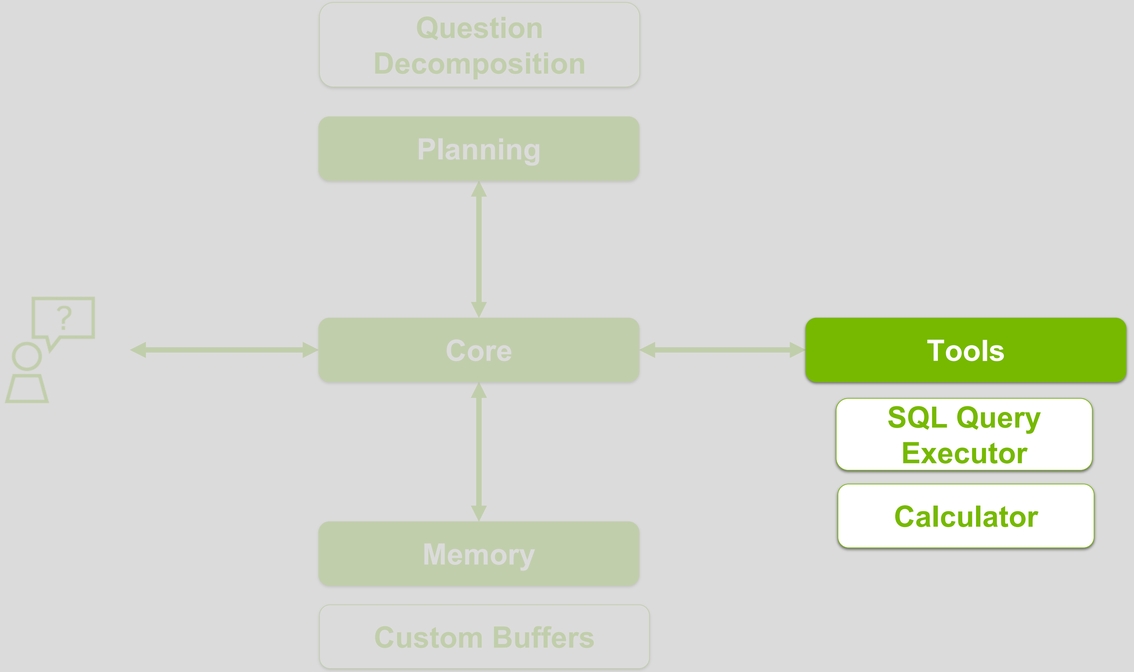
Data Agent

Use Case: Talking to a Relational DB for Inventory Management

```
inventory_data = [
    {"product_id": 1, "quantity": 150, "min_required": 30},
    {"product_id": 2, "quantity": 100, "min_required": 20},
    {"product_id": 3, "quantity": 120, "min_required": 30},
    {"product_id": 4, "quantity": 80, "min_required": 15},
    {"product_id": 5, "quantity": 200, "min_required": 40},
    {"product_id": 6, "quantity": 150, "min_required": 25},
    . . .
]
products_data = [
    {"name": "iPhone 13 Pro", "description": "Apple flagship smartphone", "price": 999.99, "supplier_id": 2},
    {"name": "iPhone SE", "description": "Apple budget smartphone", "price": 399.99, "supplier_id": 2},
    {"name": "Google Pixel 6", "description": "Google's latest smartphone", "price": 599.00, "supplier_id": 4},
    {"name": "Google Pixel 5a", "description": "Affordable Google smartphone", "price": 449.00, "supplier_id": 4},
    . . .
]
suppliers_data = [
    {"name": "Samsung Electronics", "address": "Seoul, South Korea", "contact": "800-726-7864"},
    {"name": "Apple Inc.", "address": "Cupertino, California, USA", "contact": "800-692-7753"},
    . . .
]
```

Disclaimer: - The information above are just placeholders for an example and not intended to reflect the actual details.

Data Agent - Tools



- **Query Database:** Query the Database for information.
- **Calculator:** For any basic calculations needed to do after querying the data.

```
"""<s> [INST]You are an agent capable of using a variety
of TOOLS to answer a data analytics question. Always use
MEMORY to help select the TOOLS to be used.

MEMORY
<Empty>

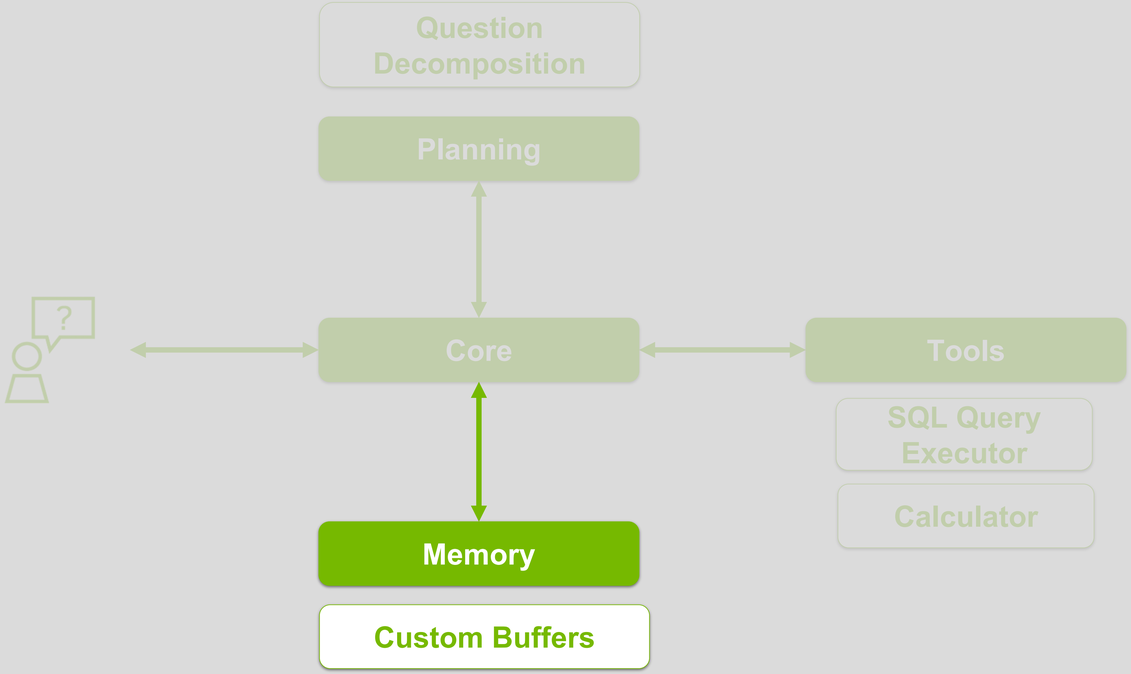
TOOLS
- Generate Final Answer: Use if answer to User's question
can be given with MEMORY
- Calculator: Use this tool to solve mathematical
problems.
- Query_Database: Write an SQL Query to query the
Database.

ANSWER FORMAT
```json
{
    "tool_name": "Calculator"
}
```
[/INST]
User: {User's Question}

Assistant: ```json
{
    "tool_name": """
```

Initial Agent Prompt

Data Agent - Memory



We will use a simple buffer/list to keep track of all the agent's actions.

```
"""<s> [INST]You are an agent capable of using a variety of TOOLS to answer a data analytics question. Always use MEMORY to help select the TOOLS to be used.

MEMORY
<Empty>

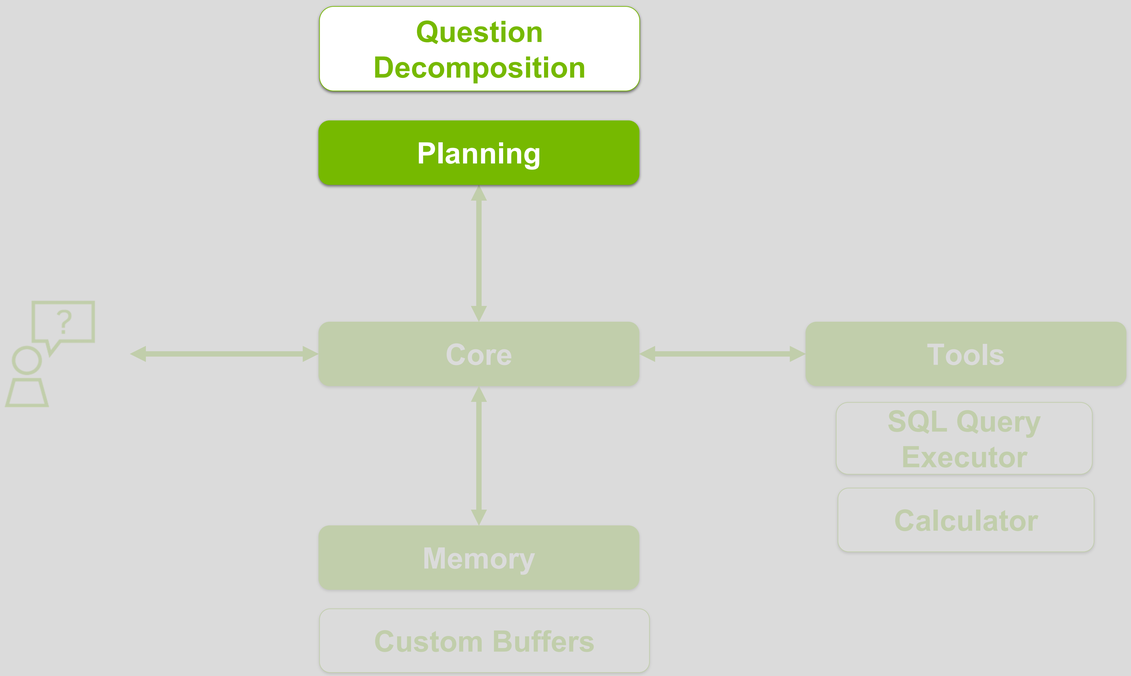
TOOLS
- Generate Final Answer: Use if answer to User's question can be given with MEMORY
- Calculator: Use this tool to solve mathematical problems.
- Query_Database: Write an SQL Query to query the Database.

ANSWER FORMAT
```json
{
    "tool_name": "Calculator"
}
```
[/INST]
User: {User's Question}

Assistant: ```json
{
    "tool_name": """
```

Initial Agent Prompt

Data Agent - Planning



A linear greedy approach. The way we achieve this is by creating a "faux tool": "Generating final answer" allowing it to iterate until the agent acquired all the relevant info to generate the final response

```
"""<s> [INST]You are an agent capable of using a variety of TOOLS to answer a data analytics question. Always use MEMORY to help select the TOOLS to be used.

MEMORY
<Empty>

TOOLS
- Generate Final Answer: Use if answer to User's question can be given with MEMORY
- Calculator: Use this tool to solve mathematical problems.
- Query_Database: Write an SQL Query to query the Database.

ANSWER FORMAT
```json
{
    "tool_name": "Calculator"
}
```
[/INST]
User: {User's Question}

Assistant: ```json
{
    "tool_name": """
```

Initial Agent Prompt

User Question: How much excess inventory do we have for Google Pixel 6?

```
def Agent(question, memory):
    tool_selection = LLM(prompt+question+memory)

    if tool_selection['tool_name'] == "Query_Database":
        Generate_SQL_Query(question, memory, schema)
        QueryDB
        Append results to memory
        Agent(question, memory)

    if tool_selection['tool_name'] == "Generate Final
Answer":
        Final_Output = llm(question, memory)
        return final output

    if tool_selection['tool_name'] == "Calculator":
        Ask_Math_Question_To_LLM(question, memory)
        Append results to memory
        Agent(question, memory)
```

Tool Selector(Query Database)

Generate SQL Query and Query Database

```
Retrieved Information: [(7, 7, 100, 20,
'Google Pixel 6')]
```

Tool Selector(Calculator)

A: 80

Gen Final Answer

A: Based on the retrieved information from the inventory system , you currently have 800 units of excess inventory for the ‘Google Pixel 6’. This calculation is derived by subtracting the minimum required quantity (20) from the current quantity in stock (100).

LLM used to showcase this example is Mixtral 8x7B Instruct

User Question: How much excess inventory do we have for Google Pixel 6?

```
def Agent(question, memory):
    tool_selection = LLM(prompt+question+memory)

    if tool_selection['tool_name'] == "Query_Database":
        Generate_SQL_Query(question, memory, schema)
        QueryDB
        Append results to memory
        Agent(question, memory)

    if tool_selection['tool_name'] == "Generate Final
Answer":
        Final_Output = llm(question, memory)
        return final output

    if tool_selection['tool_name'] == "Calculator":
        Ask_Math_Question_To_LLM(question, memory)
        Append results to memory
        Agent(question, memory)
```

Tool Selector(Query Database) →

Generate SQL Query and Query Database

```
SQL Query:
SELECT
    inventory.id,
    inventory.product_id,
    inventory.quantity,
    inventory.min_required,
    products.name
FROM
    inventory
JOIN
    products ON inventory.product_id =
products.id
WHERE
    products.name = 'Google Pixel 6';
```

```
Retrieved Information: [(7, 7, 100, 20,
'Google Pixel 6')]
```

LLM used to showcase this example is Mixtral 8x7B Instruct

NVIDIA

User Question: How much excess inventory do we have for Google Pixel 6?

```
def Agent(question, memory):
    tool_selection = LLM(prompt+question+memory)

    if tool_selection['tool_name'] == "Query_Database":
        Generate_SQL_Query(question, memory, schema)
        QueryDB
        Append results to memory
        Agent(question, memory)

    if tool_selection['tool_name'] == "Generate Final Answer":
        Final_Output = llm(question, memory)
        return final output

    if tool_selection['tool_name'] == "Calculator":
        Ask_Math_Question_To_LLM(question, memory)
        Append results to memory
        Agent(question, memory)
```

Tool Selector(Query Database) → Generate SQL Query and Query Database

```
"""
<s> [INST]You are an agent capable of using a variety of TOOLS to answer a data analytics question. Always use MEMORY to help select the TOOLS to be used.

MEMORY
Previous Question:  How much excess inventory do we have for 'Google Pixel 6'?
SQL Query:
SELECT
    inventory.id,
    inventory.product_id,
    inventory.quantity,
    inventory.min_required,
    products.name
FROM
    inventory
JOIN
    products ON inventory.product_id = products.id
WHERE
    products.name = 'Google Pixel 6';
Retrieved Information: [(7, 7, 100, 20, 'Google Pixel 6')]

TOOLS
- Generate Final Answer: Use if answer to User's question can be given with MEMORY
- Calculator: Use this tool to solve mathematical problems.
- Query_Database: Write an SQL Query to query the Database.

ANSWER FORMAT
```json
```

Updated Prompt

LLM used to showcase this example is Mixtral 8x7B Instruct

NVIDIA
```

User Question: How much excess inventory do we have for Google Pixel 6?

Tool Selector(Query Database)

Generate SQL Query and Query Database

```
Retrieved Information: [(7, 7, 100, 20,
'Google Pixel 6')]
```

Tool Selector(Calculator)

A: 80

```
def Agent(question, memory):
    tool_selection = LLM(prompt+question+memory)

    if tool_selection['tool_name'] == "Query_Database":
        Generate_SQL_Query(question, memory, schema)
        QueryDB
        Append results to memory
        Agent(question, memory)

    if tool_selection['tool_name'] == "Generate Final
Answer":
        Final_Output = llm(question, memory)
        return final output

    if tool_selection['tool_name'] == "Calculator":
        Ask_Math_Question_To_LLM(question, memory)
        Append results to memory
        Agent(question, memory)
```

LLM used to showcase this example is Mixtral 8x7B Instruct

User Question: How much excess inventory do we have for Google Pixel 6?

```
def Agent(question, memory):
    tool_selection = LLM(prompt+question+memory)

    if tool_selection['tool_name'] == "Query_Database":
        Generate_SQL_Query(question, memory, schema)
        QueryDB
        Append results to memory
        Agent(question, memory)

    if tool_selection['tool_name'] == "Generate Final Answer":
        Final_Output = llm(question, memory)
        return final output

    if tool_selection['tool_name'] == "Calculator":
        Ask_Math_Question_To_LLM(question, memory)
        Append results to memory
        Agent(question, memory)
```

Tool Selector(Query Database)

Generate SQL Query and Query Database

```
Retrieved Information: [(7, 7, 100, 20, 'Google Pixel 6')]
```

Tool Selector(Calculator)

A: 80

Gen Final Answer

A: Based on the retrieved information from the inventory system , you currently have 800 units of excess inventory for the ‘Google Pixel 6’. This calculation is derived by subtracting the minimum required quantity (20) from the current quantity in stock (100).

LLM used to showcase this example is Mixtral 8x7B Instruct

NVIDIA

API Agent

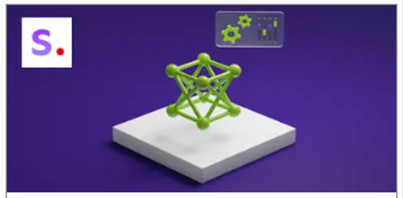
Use Case: Chain Models APIs to make a “brainstorming copilot”.

Mixtral 8x7B Instruct

Text Generation

Mixtral 8x7B Instruct is a language model that can follow instructions, complete requests, and generate creative text...

View Labels | Learn More



Stable Diffusion XL

Image Generation

Stable Diffusion XL (SDXL) enables you to generate expressive images with shorter prompts and insert words inside images.

View Labels | Learn More

```
class TextGenerator:
    def __init__(self): ...

    def generate(self, prompt, temperature=0.01, max_tokens=1024): ...

class CodeGenerator:
    def __init__(self): ...

    def generate(self, prompt, temperature=0.2, max_tokens=1024): ...

class ImageGenerator:
    def __init__(self): ...

    def generate(self, prompt, negative_prompt): ...
```

Code Llama 13B

Code Generation

Code Llama is an LLM capable of generating code, and natural language about code, from both code and natural...

View Labels | Learn More

User Question: Write a tweet for my ad campaign around selling more detergent, the name of the product is WishyWash, with a new UltraClean formula priced at $ 4.99. Generate an image to go with it. Actually, while you are at it add that the new product has a softner in the tweet. Brainstorm some other ideas for marketing apart from the tweet.

Plan

```
{
    "function": "ImageGenerator",
    "arguments": {
        "prompt": "A bright and clean laundry room with a large bottle of WishyWash detergent, featuring the new UltraClean formula and softener, placed prominently on a shelf.",
        "negative_prompt": "No clutter, no other brands, only WishyWash."
    }
},
{
    "function": "TextGenerator",
    "arguments": {
        "prompt": "Compose a tweet to promote the new WishyWash detergent with the UltraClean formula and softener at $4.99. Make sure to highlight its benefits and competitive pricing."
    }
},
{
    "function": "TextGenerator",
    "arguments": {
        "prompt": "Generate ideas for marketing campaigns to increase WishyWash detergent sales, specifically focusing on the new UltraClean formula and softener."
    }
}
```

User Question: Write a tweet for my ad campaign around selling more detergent, the name of the product is WishyWash, with a new UltraClean formula priced at $ 4.99. Generate an image to go with it. Actually, while you are at it add that the new product has a softner in the tweet. Brainstorm some other ideas for marketing apart from the tweet.

Plan


```
{
    "function": "ImageGenerator",
    "arguments": {
        "prompt": "A bright and clean laundry room with a large bottle of WishyWash detergent, featuring the new UltraClean formula and softener, placed prominently on a shelf.",
        "negative_prompt": "No clutter, no other brands, only WishyWash."
    }
},
{
    "function": "TextGenerator",
    "arguments": {
        "prompt": "Compose a tweet to promote the new WishyWash detergent with the UltraClean formula and softener at $4.99. Make sure to highlight its benefits and competitive pricing."
    }
},
{
    "function": "TextGenerator",
    "arguments": {
        "prompt": "Generate ideas for marketing campaigns to increase WishyWash detergent sales, specifically focusing on the new UltraClean formula and softener."
    }
}
```


Execute


```
{    "function": "ImageGenerator",
    "arguments": {
        "prompt": "A bright and clean laundry room with a large bottle of WishyWash detergent, featuring the new UltraClean formula and softener, placed prominently on a shelf.",
        "negative_prompt": "No clutter, no other brands, only WishyWash."
    }
}
```


Gen Image

User Question: Write a tweet for my ad campaign around selling more detergent, the name of the product is WishyWash, with a new UltraClean formula priced at $ 4.99. Generate an image to go with it. Actually, while you are at it add that the new product has a softner in the tweet. Brainstorm some other ideas for marketing apart from the tweet.

Plan

```
{
    "function": "ImageGenerator",
    "arguments": {
        "prompt": "A bright and clean laundry room with a large bottle of WishyWash detergent, featuring the new UltraClean formula and softener, placed prominently on a shelf.",
        "negative_prompt": "No clutter, no other brands, only WishyWash."
    }
},
{
    "function": "TextGenerator",
    "arguments": {
        "prompt": "Compose a tweet to promote the new WishyWash detergent with the UltraClean formula and softener at $4.99. Make sure to highlight its benefits and competitive pricing."
    }
}
```

Execute

Gen Tweet

🎉 Exciting news, folks! 🎉 Introducing WishyWash's UltraClean formula & softener in one! 🧺🍋 Say goodbye to stubborn stains and hello to irresistibly soft clothes. 🌟 Now at an unbeatable price of $4.99! 💸💥 Shop now for a cleaner, fresher home! 🏡🛒 #WishyWash #UltraClean #AffordableLuxury

```
    }
},
{
    "function": "TextGenerator",
    "arguments": {
        "prompt": "Generate ideas for marketing campaigns to increase WishyWash detergent sales, specifically focusing on the new UltraClean formula and softener."
    }
}
```

User Question: Write a tweet for my ad campaign around selling more detergent, the name of the product is WishyWash, with a new UltraClean formula priced at $ 4.99. Generate an image to go with it. Actually, while you are at it add that the new product has a softner in the tweet. Brainstorm some other ideas for marketing apart from the tweet.

Plan

```
{
    "function": "ImageGenerator",
    "arguments": {
        "prompt": "A bright and clean laundry room with a
large bottle of WishyWash detergent, featuring the new
UltraClean formula and softener, placed prominently on a
shelf.",
_prompt": "No clutter, no other brands,
TextGenerator",
{
"Compose a tweet to promote the new
nt with the UltraClean formula and
. Make sure to highlight its benefits and
ng."
TextGenerator",
{
"Generate ideas for marketing campaigns
Wash detergent sales, specifically
ew UltraClean formula and softener."
```

Execute

```
{
    "function": "TextGenerator",
    "arguments": {
        "prompt": "Generate ideas for
marketing campaigns to increase
WishyWash detergent sales, specifically
focusing on the new UltraClean formula
and softener."
    }
}
```

Gen Ideas

1. **Educational Content**: Share blog posts, social media updates, or videos explaining the benefits of the UltraClean formula and softener. Discuss how it effectively removes stains, leaves clothes soft, and is gentle on fabrics.

2. **Testimonials and Reviews:** Encourage satisfied customers to share their experiences using WishyWash UltraClean on your website and social media platforms. Positive reviews and testimonials can significantly influence potential buyers.

3. **Comparative Advertising:** Create ads that compare WishyWash UltraClean to competitors, highlighting the unique advantages of your product, such as superior cleaning power, gentleness on clothes, and affordability.

4. **Limited-Time Offers:** Promote special discounts or bundle deals for a limited time to create a sense of urgency. For example, offer a discount when customers buy both the UltraClean formula and softener together.

5. **Influencer Partnerships:** Collaborate with influencers in the home and lifestyle space to promote WishyWash UltraClean and softener. Influencers can create content showing how they use the product, share their honest opinions, and reach a wider audience.

6. **Interactive Content:** Develop quizzes, polls, or games on your website or social media platforms to engage your audience and create awareness about the new UltraClean formula and softener.

7. **Sustainability Initiatives:** Highlight any eco-friendly aspects of your product, such as biodegradable packaging or sustainable production methods, to appeal to environmentally conscious consumers.

8. **Before-and-After Visuals:** Share images or videos demonstrating the cleaning power of WishyWash UltraClean. Visuals can be a powerful tool to showcase the effectiveness of your product.

9. **Social Responsibility:** Share stories about how WishyWash is involved in the community or supports good causes. Consumers are more likely to support businesses that align with their values.

10. **Email Marketing:** Send personalized emails to your mailing list, introducing the new UltraClean formula and softener, offering exclusive discounts, and sharing educational content.

Designing a robust Agent

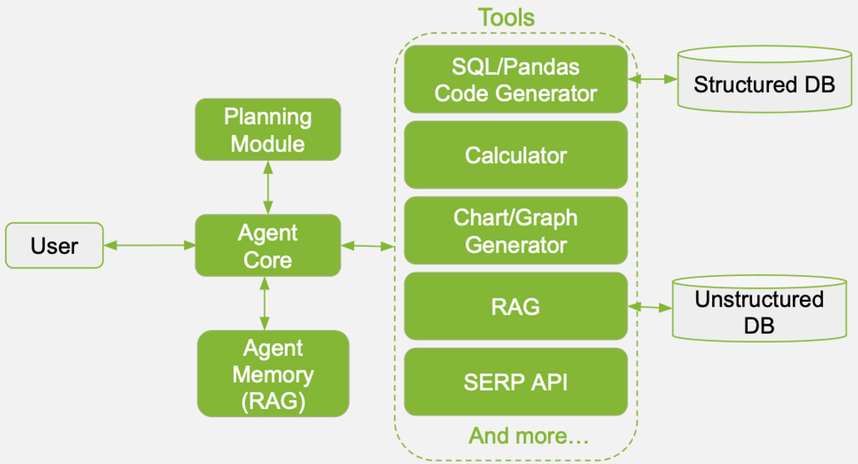
Scaling the Tools

Working with Multiple Vector Databases

Better Planning Module

Say Instead of 3 tables, 3 tools or 3 APIs, we have 100k + tables and 100 tools. How do we accommodate this?

- Use RAG to retrieve top 5 most relevant tools.
- Can Apply to Memory, Database Schema, or any other "options" that the Agent needs to consider.



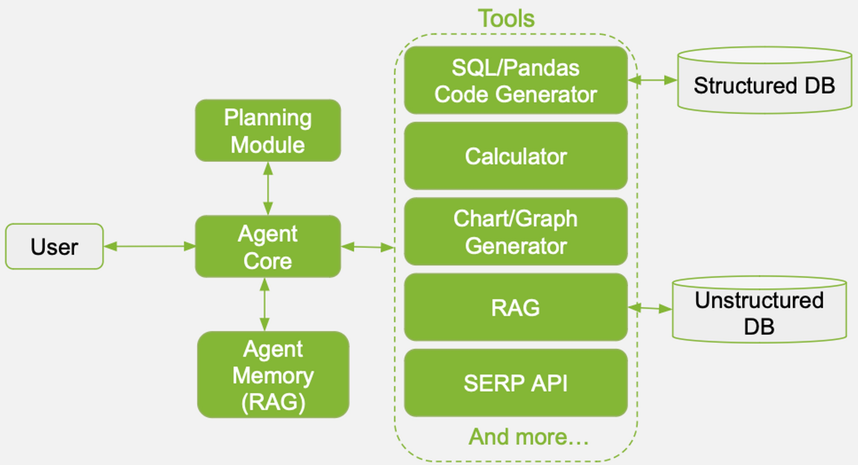
Designing a robust Agent

Scaling the Tools

Working with Multiple Vector Databases

Better Planning Module

Topical router to direct the queries to the correct database.



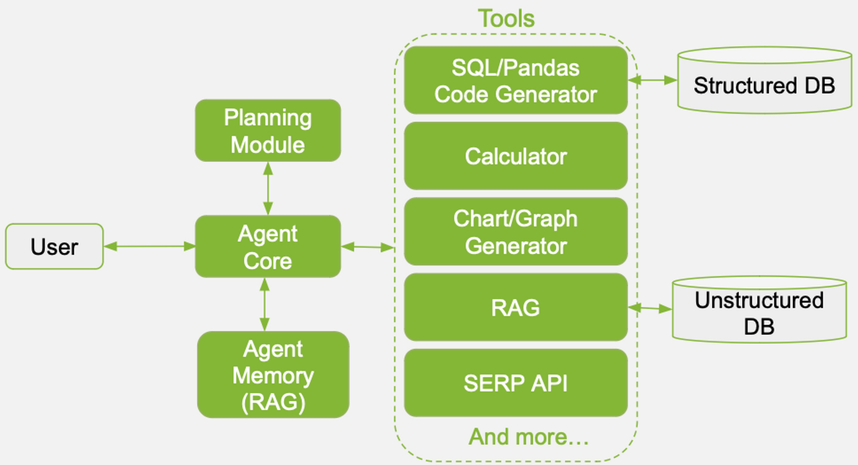
Designing a robust Agent

Scaling the Tools

Working with Multiple Vector Databases

Better Planning Module

Task Decomposition Module or a Plan compiler of sorts to generate a more efficient plan of execution.



Multimodal data in the wild!

NVIDIA DEVELOPER Home Blog Forums Docs Downloads Training

Technical Blog Search blog Filter

Data Center / Cloud English

Setting New Records in MLPerf Inference v3.0 with Full-Stack Optimizations for AI

Apr 05, 2023 +9 Like Discuss (0)

By Ashraf Eassa, Ashwin Nanjappa, Zhihan Jiang, Jesus Corbal San Adrian, Jinho Suh, Yiheng Zhang and Nikolay Markovskiy

The most exciting computing applications currently rely on training and running inference on complex AI models, often in demanding, real-time deployment scenarios. High-performance, accelerated AI platforms are needed to meet the demands of these applications and deliver the best user experiences.

New AI models are constantly being invented to enable new capabilities, and AI-powered applications often rely on many such models working in tandem. This means that an AI platform must be able to run the widest range of workloads and deliver excellent performance on all of them. MLPerf Inference—now in its seventh edition with v3.0—is a trusted, peer-reviewed suite of standardized inference performance tests that represents many such AI models.

AI applications run everywhere, from the largest hyperscale data centers to compact edge devices. MLPerf Inference represents both data center and edge environments. It also represents a range of real-world scenarios, such as offline (batch) processing, and latency-constrained server, single-stream, and multi-stream scenarios. This balance of workload breadth and depth ensures that MLPerf Inference is a valuable resource for those looking to select AI infrastructure to best meet the needs of a diverse set of deployments.

In the MLPerf Inference v3.0 round, NVIDIA submitted results on several products, including NVIDIA H100 Tensor Core GPUs (both SXM and PCIe add-in-card form factors) based on the NVIDIA Hopper architecture, the recently announced NVIDIA L4 Tensor Core GPU powered by the NVIDIA Ada Lovelace GPU architecture, as well as the NVIDIA Jetson AGX Orin and NVIDIA

NVIDIA AI and H100 Tensor Core GPU deliver record results

NVIDIA H100 Tensor Core GPUs, which made their MLPerf Training debut just 6 months ago, set new per-accelerator performance records across all MLPerf Training v3.0 workloads. Looking at the NVIDIA single-node DGX H100 results this round, performance increased by up to 17% in just 6 months on the same hardware through software improvements alone. Compared to the NVIDIA A100 Tensor Core GPU submission in MLPerf Training v2.1, the latest H100 submission delivered up to 3.1x more performance per accelerator.

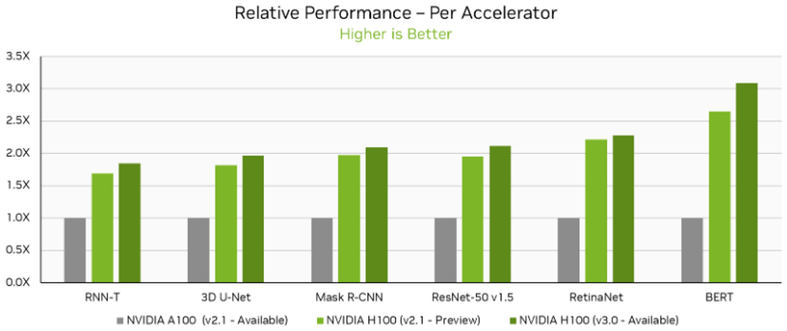


Figure 1. Relative per-accelerator performance of NVIDIA A100 in MLPerf Training v2.1, and NVIDIA H100 in MLPerf Training v2.1 and v3.0.

MLPerf result IDs: 2.1-2091, 2.1-2062, 2.1-2059, 2.1-2061, 2.1-2060, 3.0-2062, 3.0-2064. The MLPerf name and logo are trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

In this round, NVIDIA submitted results using the NVIDIA "Pre-Eos" AI supercomputer on up to 768 H100 GPUs. NVIDIA also made a joint submission with cloud service provider CoreWeave, using up to 3,584 H100 GPUs with the CoreWeave publicly available NVIDIA HGX H100 infrastructure.

Across these submissions, the NVIDIA AI platform with H100 GPUs set new time-to-train records at scale across every workload, including the new LLM workload.

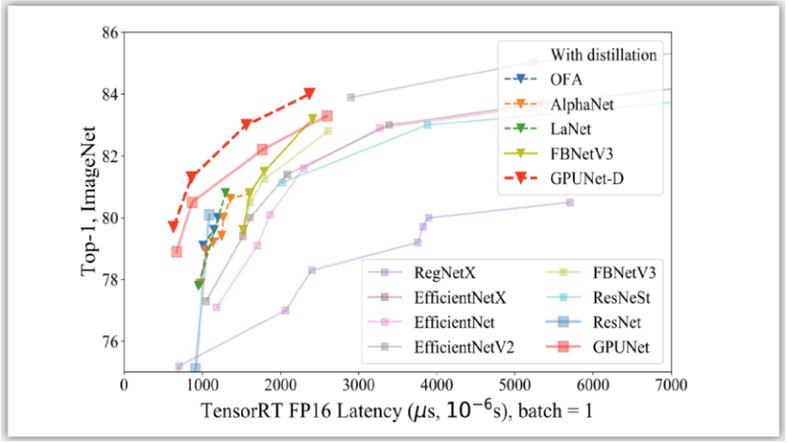
Benchmark	**Max Scale Records**(minutes)
Large language model (GPT-3)	10.9
Natural language processing (BERT)	0.13 (8 seconds)
Recommendation (DLRMv2)	1.61
Object detection, heavyweight (Mask R-CNN)	1.47
Object detection, lightweight (RetinaNet)	1.51

Computer Vision / Video Analytics English

Beating SOTA Inference Performance on NVIDIA GPUs with GPUNet

Aug 30, 2022 +2 Like Discuss (0)

By Satish Salian, Carl (Izzy) Putterman, Linnan Wang and Krzysztof Kudrynski



Crafted by AI for AI, GPUNet is a class of convolutional neural networks designed to maximize the performance of NVIDIA GPUs using NVIDIA TensorRT.

Built using novel neural architecture search (NAS) methods, GPUNet demonstrates state-of-the-art inference performance up to 2x faster than EfficientNet-X and FBNet-V3.

The NAS methodology helps build GPUNet for a wide range of applications such that deep learning engineers can directly deploy these neural networks depending on the relative accuracy and latency targets.

GPUNet NAS design methodology

Efficient architecture search and deployment-ready models are the key goals of the NAS design methodology. This means little to no interaction with the domain experts and efficient use of cluster nodes for training potential architecture candidates. Most important is that the generated models are deployment-ready.

Crafted by AI

Multimodal RAG - What can it do?

NVIDIA AI and H100 Tensor Core GPU deliver record results

NVIDIA H100 Tensor Core GPUs, which made their MLPerf Training debut just 6 months ago, set new per-accelerator performance records across all MLPerf Training v3.0 workloads. Looking at the NVIDIA single-node DGX H100 results this round, performance increased by up to 17% in just 6 months on the same hardware through software improvements alone. Compared to the NVIDIA A100 Tensor Core GPU submission in MLPerf Training v2.1, the latest H100 submission delivered up to 3.1x more performance per accelerator.

What deep learning models were used to benchmark the relative performance of H100?

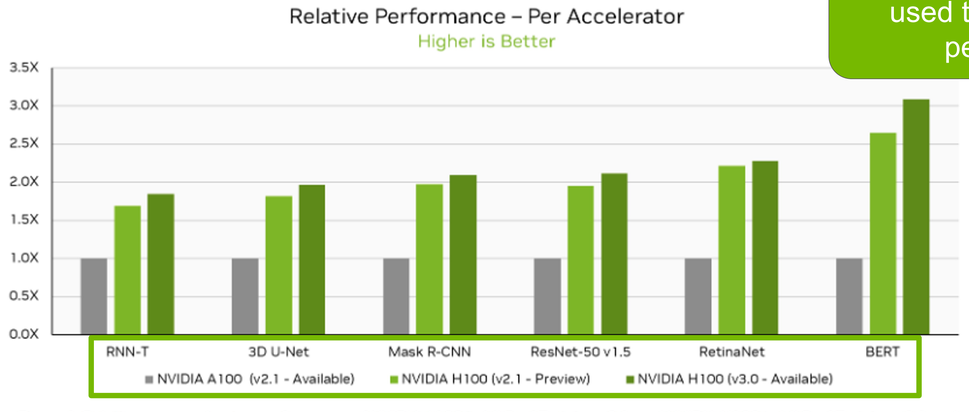


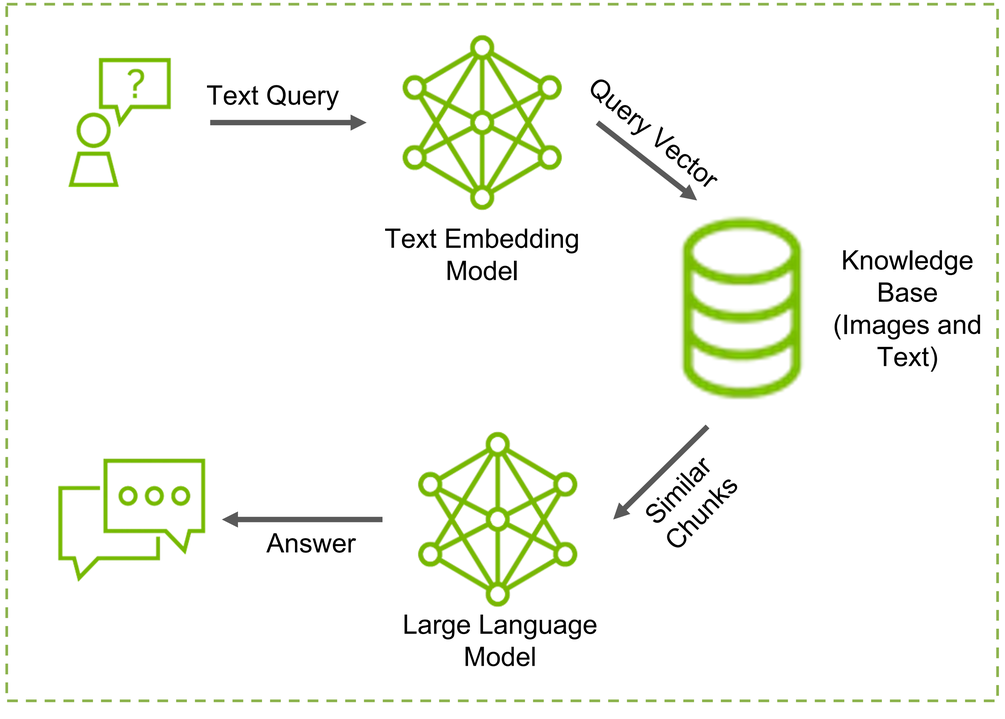
Figure 1. Relative per-accelerator performance of NVIDIA A100 in MLPerf Training v2.1, and NVIDIA H100 in MLPerf Training v2.1 and v3.0.

MLPerf result IDs: 2.1-2091, 2.1-2062, 2.1-2059, 2.1-2061, 2.1-2060, 3.0-2062, 3.0-2064. The MLPerf name and logo are trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

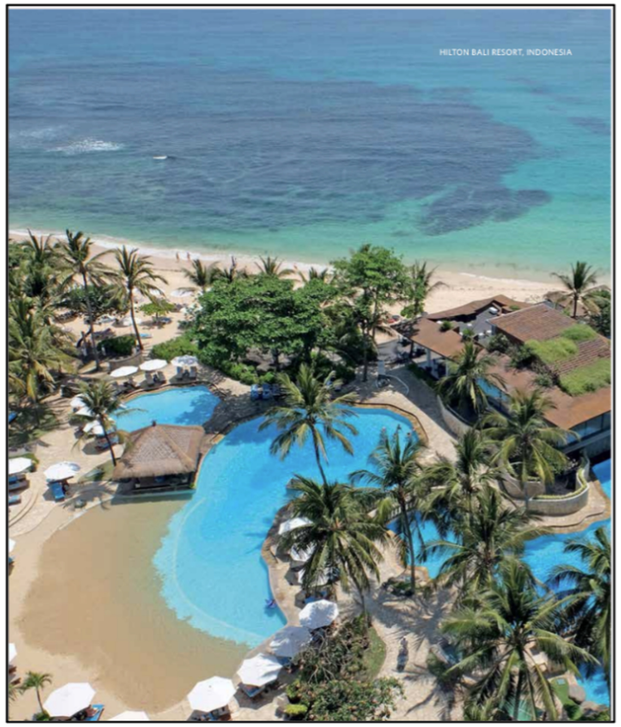
In this round, NVIDIA submitted results using the NVIDIA "Pre-Eos" AI supercomputer on up to 768 H100 GPUs. NVIDIA also made a joint submission with cloud service provider CoreWeave, using up to 3,584 H100 GPUs with the CoreWeave publicly available NVIDIA HGX H100 infrastructure.

Across these submissions, the NVIDIA AI platform with H100 GPUs set new time-to-train records at scale across every workload, including the new LLM workload.

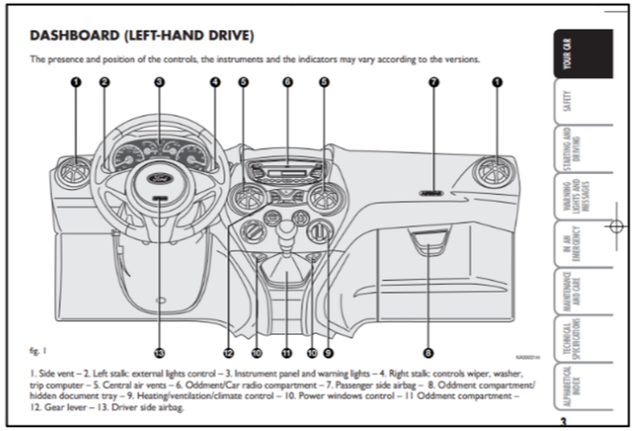
Benchmark	**Max Scale Records**(minutes)
Large language model (GPT-3)	10.9
Natural language processing (BERT)	0.13 (8 seconds)
Recommendation (DLRMv2)	1.61
Object detection, heavyweight (Mask R-CNN)	1.47
Object detection, lightweight (RetinaNet)	1.51



Why is multimodality hard?



- Hard to capture all information in text
- Very few “key points” of attention. More Focus on “General Imagery”



DASHBOARD (LEFT-HAND DRIVE)

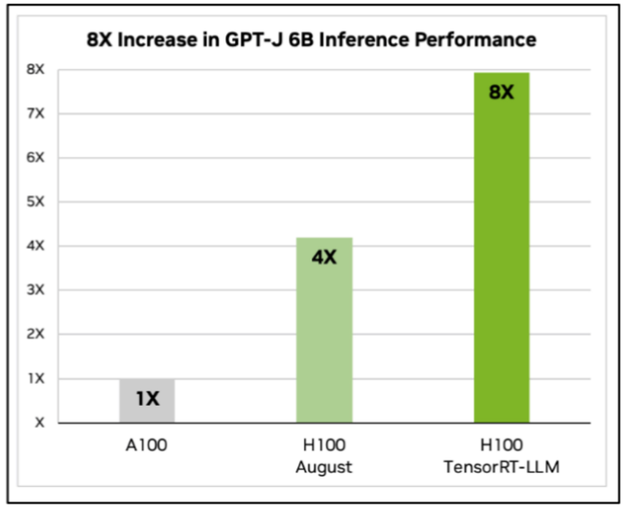
The presence and position of the controls, the instruments and the indicators may vary according to the versions.

fig. 1

1. Side vent – 2. Left stalk: external lights control – 3. Instrument panel and warning lights – 4. Right stalk: controls wiper, washer, trip computer – 5. Central air vents – 6. Oddment/Car radio compartment – 7. Passenger side airbag – 8. Oddment compartment/ hidden document tray – 9. Heating/ventilation/climate control – 10. Power windows control – 11 Oddment compartment – 12. Gear lever – 13. Driver side airbag.

YOUR CAR
SAFETY
STARTING AND DRIVING
WARNING LIGHTS AND MESSAGES
IN AN EMERGENCY
MAINTENANCE AND CARE
TECHNICAL SPECIFICATIONS
ALPHABETICAL INDEX

- Some details can be captured in Text
- Has both “key points” of attention and “General Imagery”

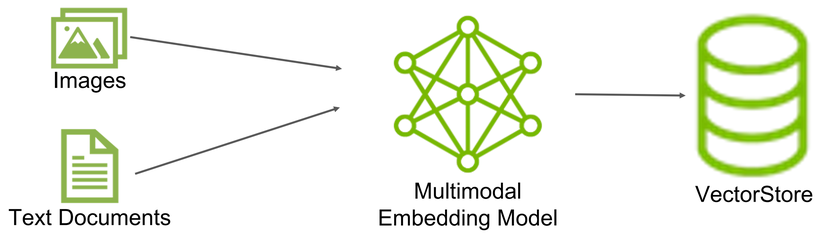


- Details can be perfectly captured in text.
- “Has key points of attention”

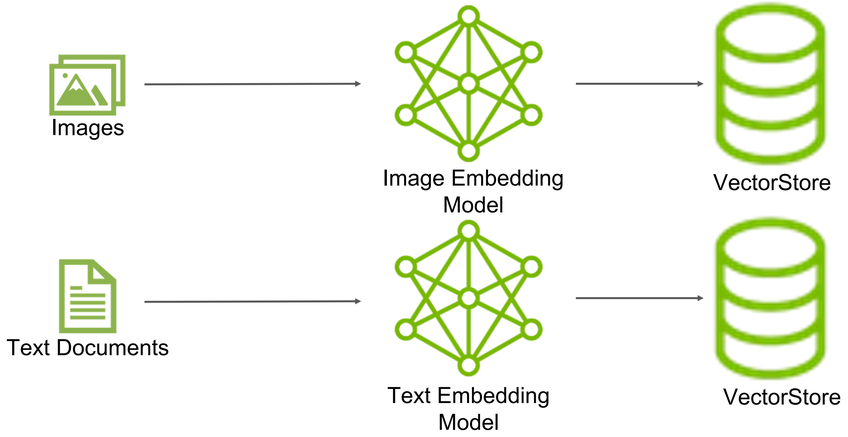
Left to Right - Increasing degree of “Concisely express-able Information”

Multimodal RAG - Approaches for Multimodal Retrieval

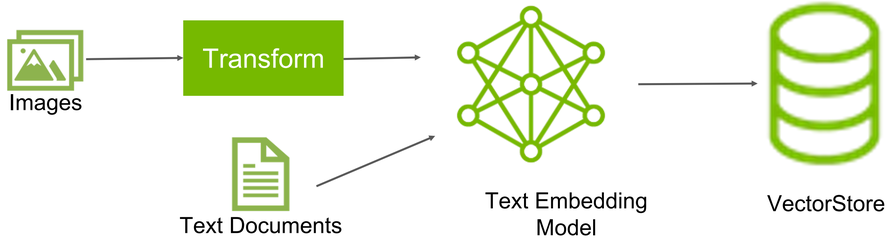
Transform all modalities into a single Vector Space



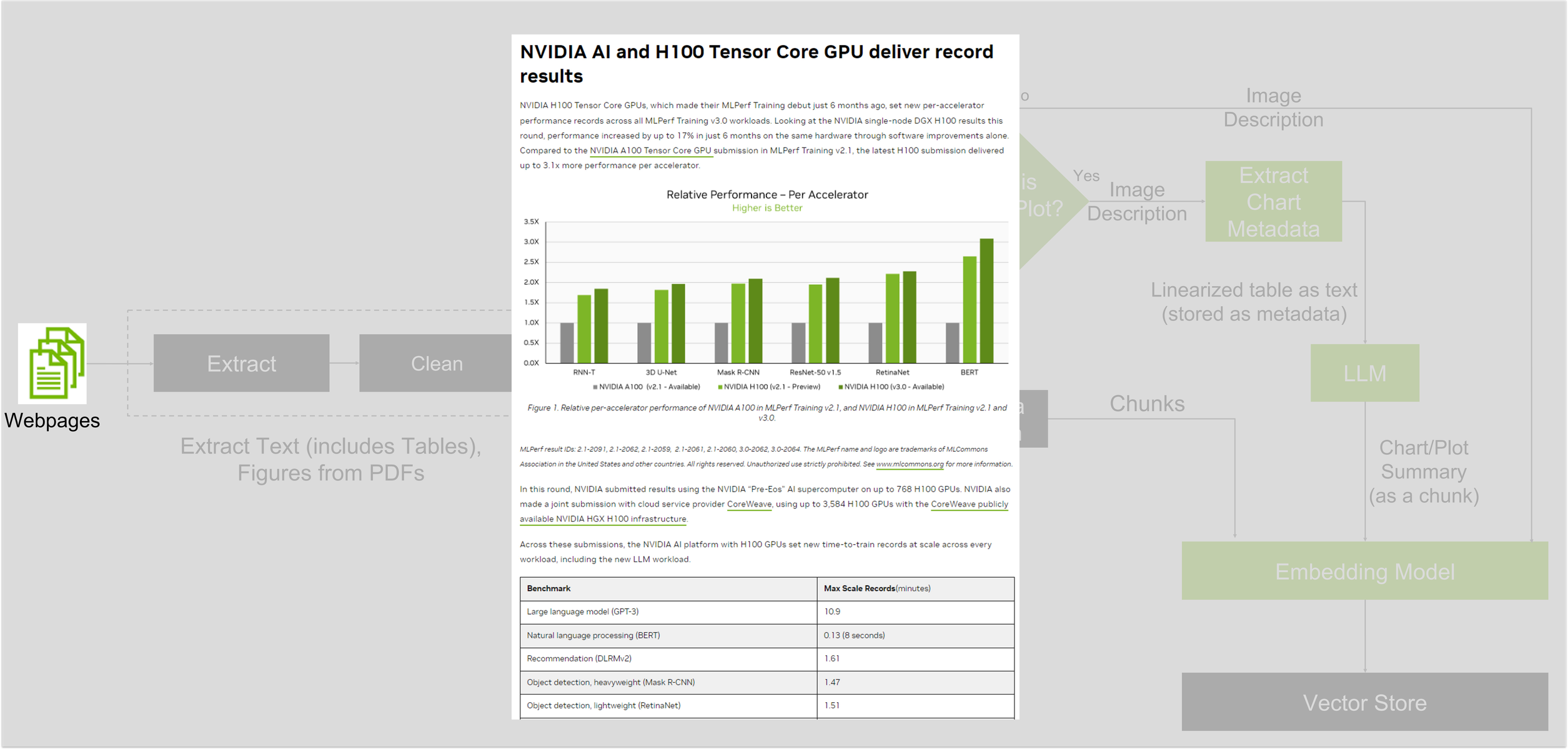
Different Modalities in Different VectorStores



Grounding all modalities into one



Multimodal RAG - Preprocessing Workflow



NVIDIA AI and H100 Tensor Core GPU deliver record results

NVIDIA H100 Tensor Core GPUs, which made their MLPerf Training debut just 6 months ago, set new per-accelerator performance records across all MLPerf Training v3.0 workloads. Looking at the NVIDIA single-node DGX H100 results this round, performance increased by up to 17% in just 6 months on the same hardware through software improvements alone. Compared to the NVIDIA A100 Tensor Core GPU submission in MLPerf Training v2.1, the latest H100 submission delivered up to 3.1x more performance per accelerator.

Figure 1. Relative per-accelerator performance of NVIDIA A100 in MLPerf Training v2.1, and NVIDIA H100 in MLPerf Training v2.1 and v3.0.

MLPerf result IDs: 2.1-2091, 2.1-2062, 2.1-2059, 2.1-2061, 2.1-2060, 3.0-2062, 3.0-2064. The MLPerf name and logo are trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

In this round, NVIDIA submitted results using the NVIDIA "Pre-Eos" AI supercomputer on up to 768 H100 GPUs. NVIDIA also made a joint submission with cloud service provider CoreWeave, using up to 3,584 H100 GPUs with the CoreWeave publicly available NVIDIA HGX H100 infrastructure.

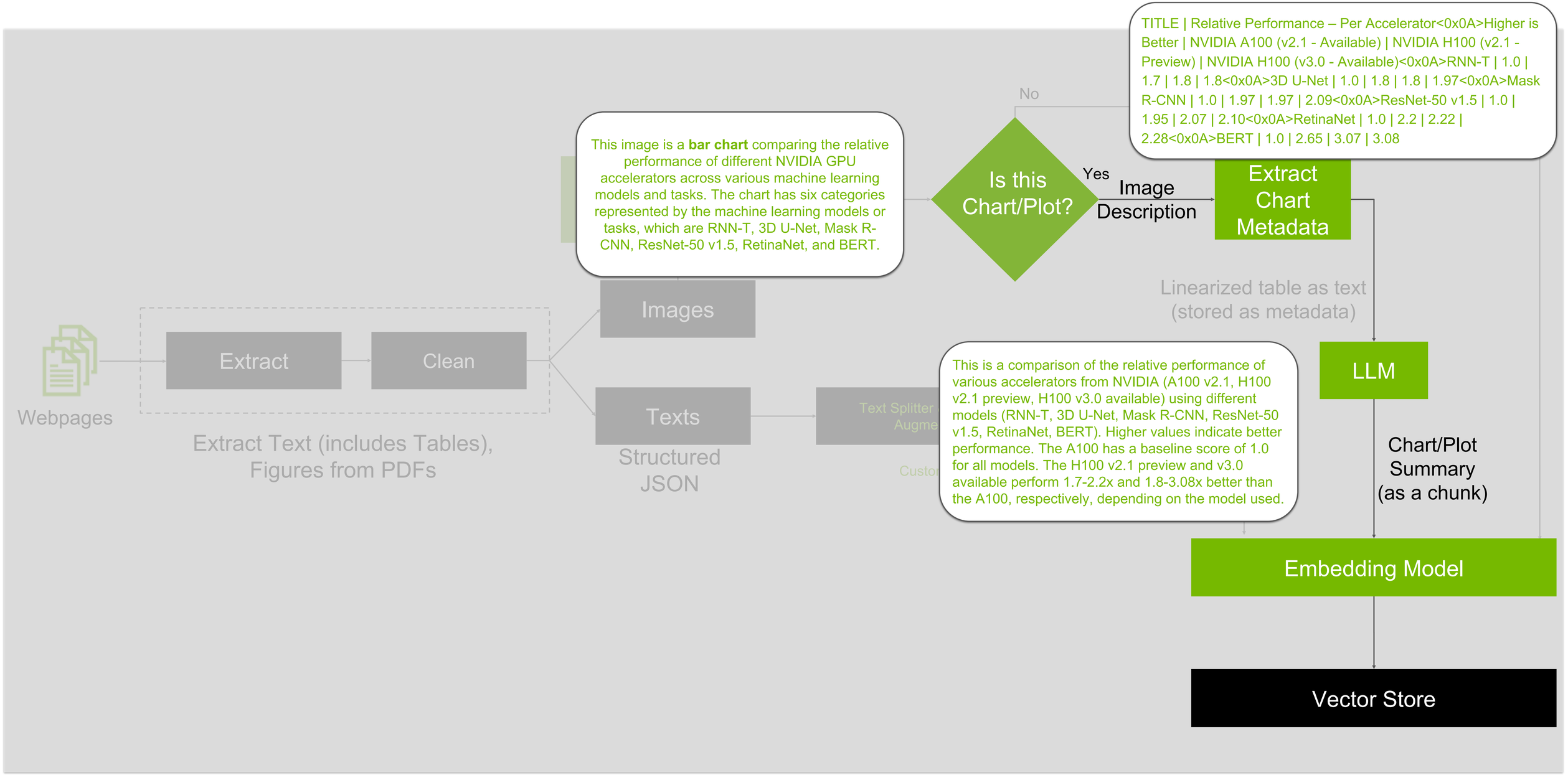
Across these submissions, the NVIDIA AI platform with H100 GPUs set new time-to-train records at scale across every workload, including the new LLM workload.

Benchmark	**Max Scale Records**(minutes)
Large language model (GPT-3)	10.9
Natural language processing (BERT)	0.13 (8 seconds)
Recommendation (DLRMv2)	1.61
Object detection, heavyweight (Mask R-CNN)	1.47
Object detection, lightweight (RetinaNet)	1.51

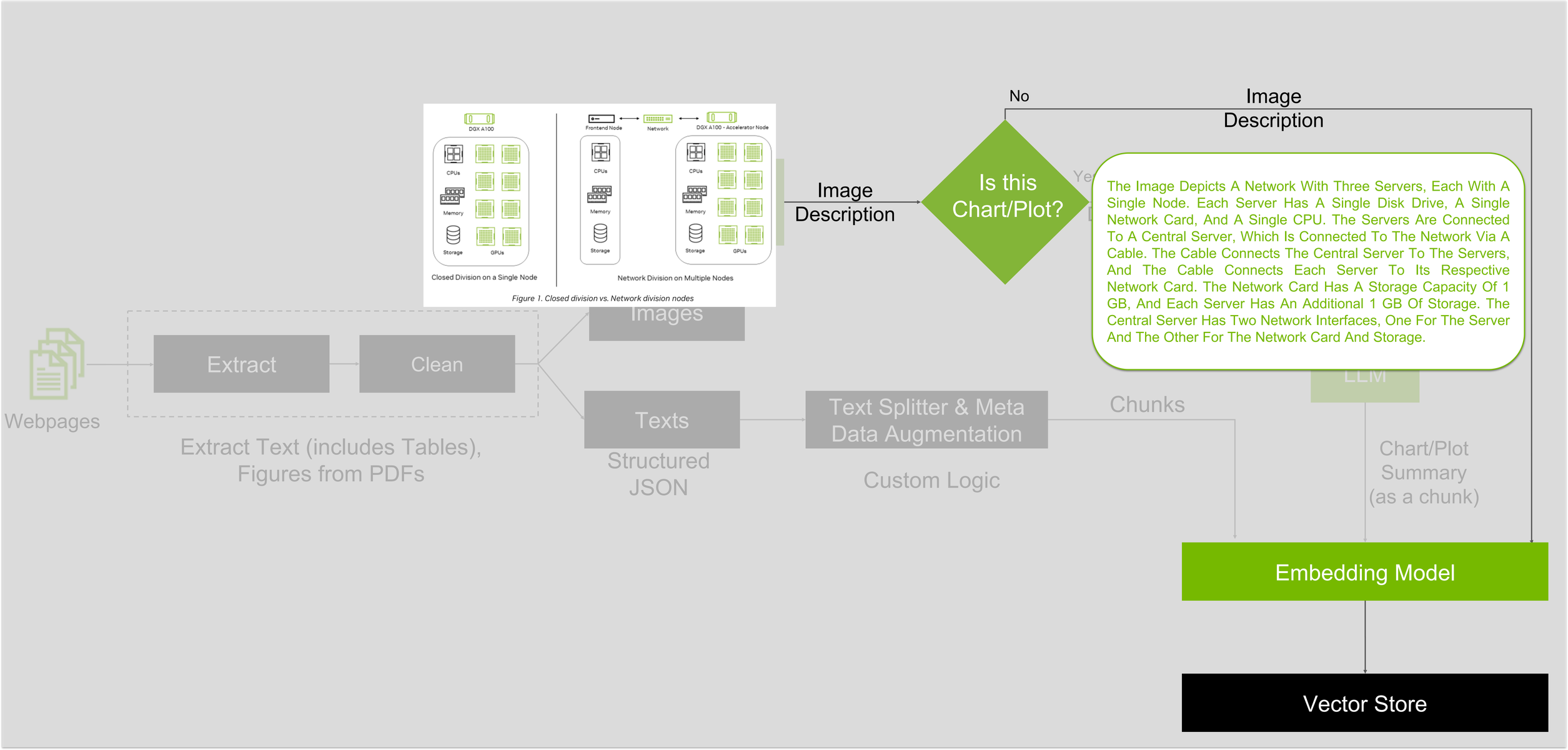
Multimodal RAG - Preprocessing Workflow

Multimodal RAG - Preprocessing Workflow

Multimodal RAG - Preprocessing Workflow

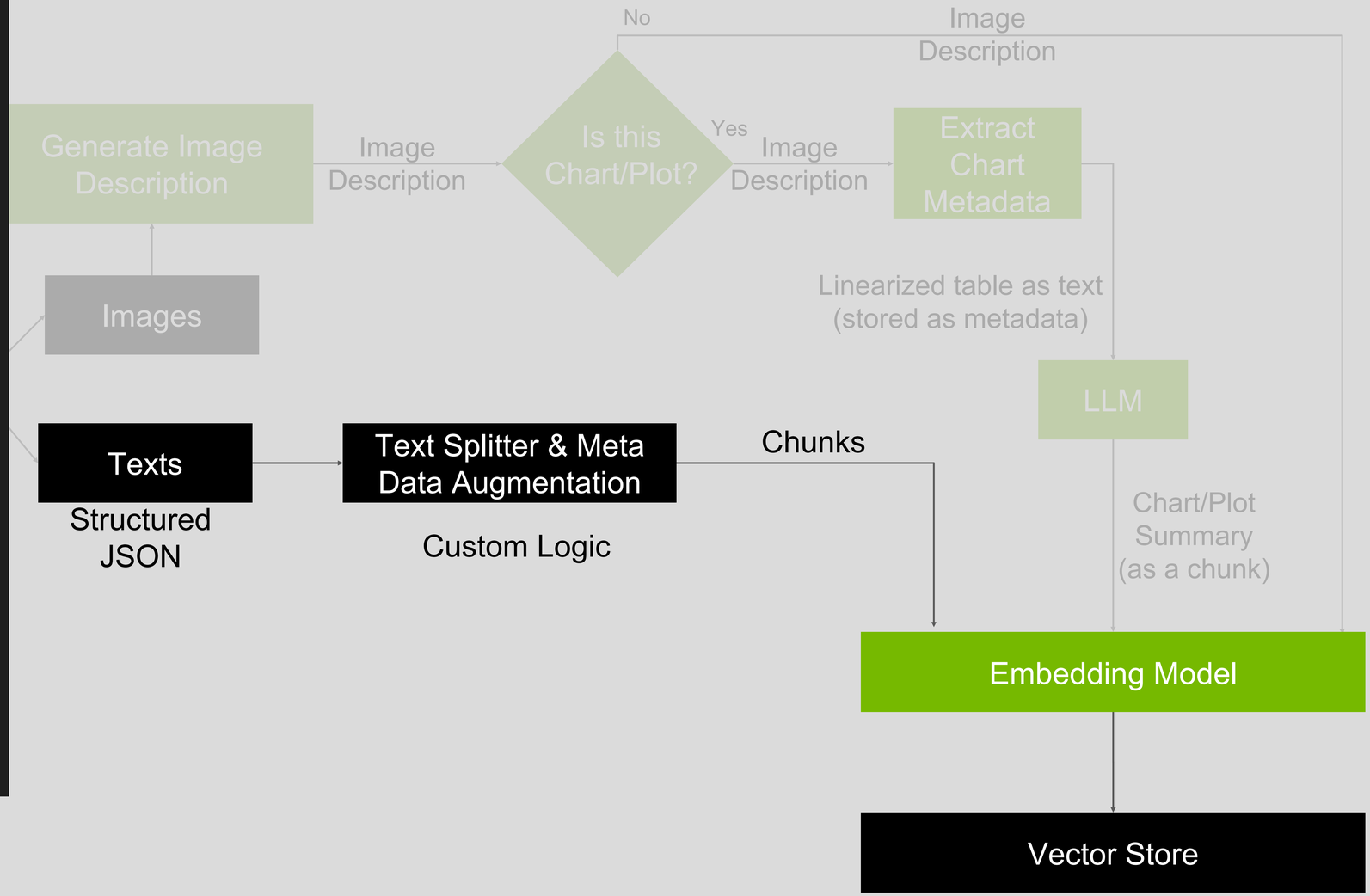


Multimodal RAG - Preprocessing Workflow

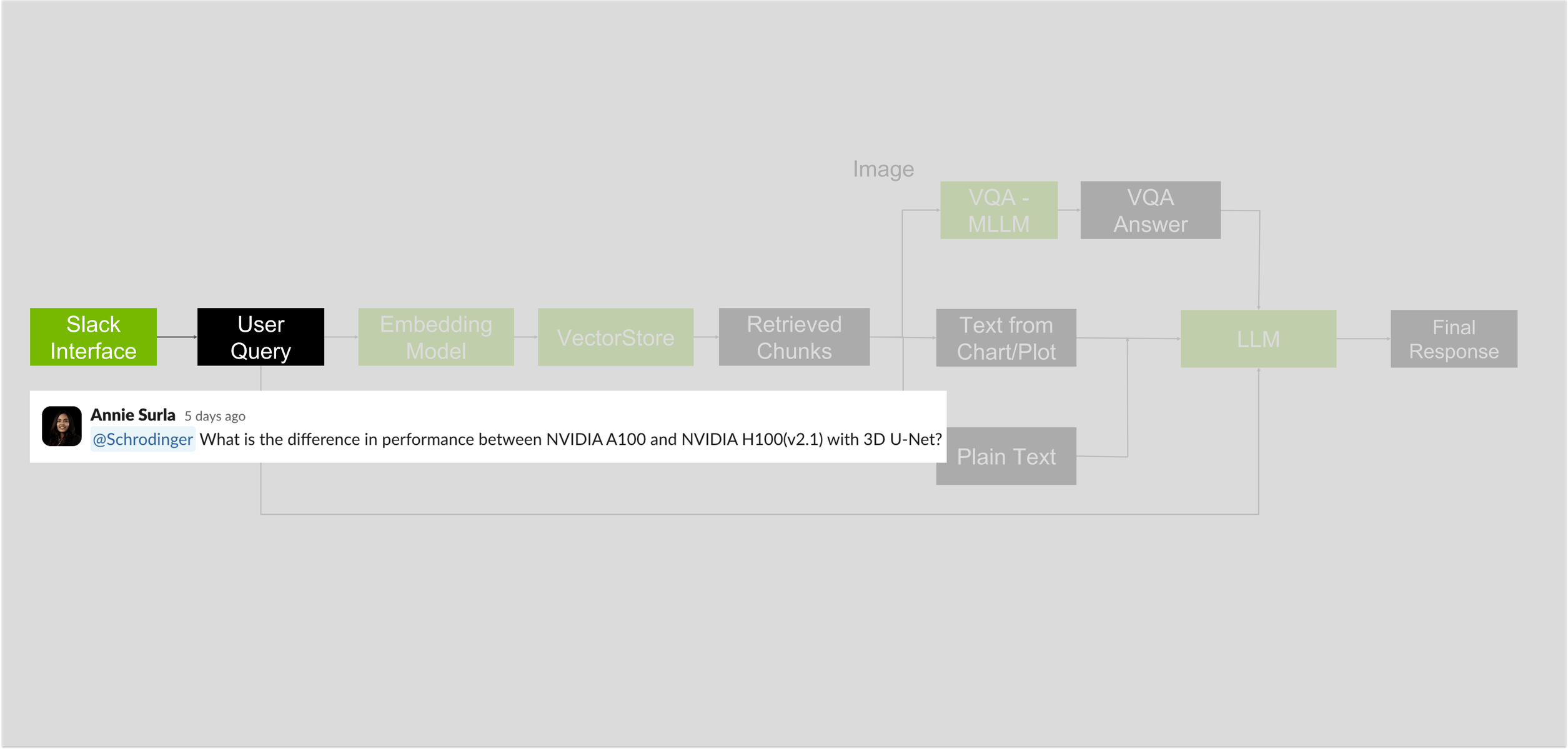


Multimodal RAG - Preprocessing Workflow

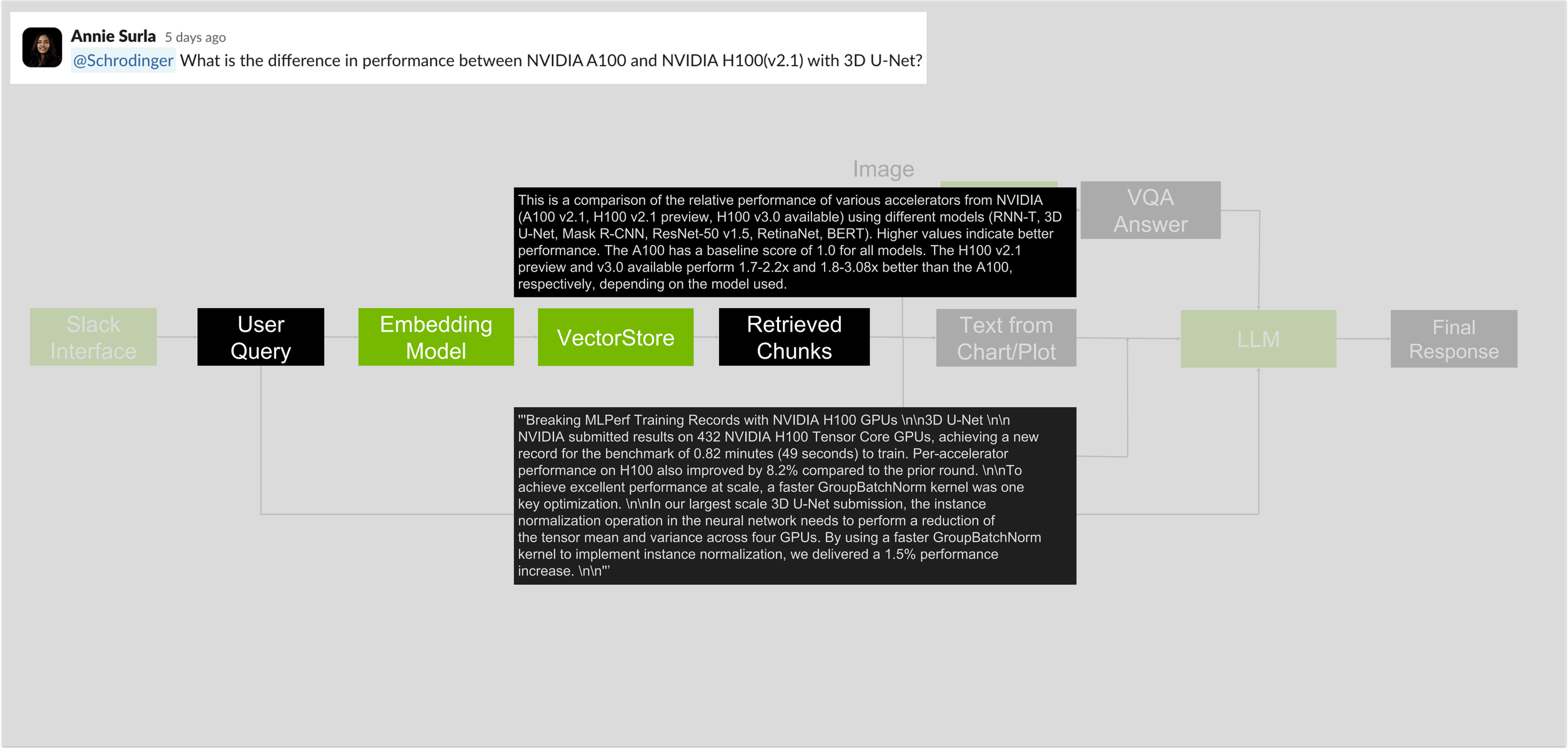

```
'''Setting New Records in MLPerf Inference v3.0 with Full-Stack
Optimizations for AI \n\nNVIDIA L4 Tensor Core GPU vaults ahead
\n\nNVIDIA L4 \n\nTensor Core GPU\n\nIn MLPerf Inference v3.0, NVIDIA
made the debut submission of the . Based on the new NVIDIA Ada
Lovelace architecture, L4 is the successor to the popular NVIDIA T4
Tensor Core GPU, delivering significant improvements for AI, video,
and graphics in the same single-slot, low-profile PCIe form factor.
\n\nNVIDIA Ada Lovelace architecture incorporates 4th Gen Tensor
Cores with FP8, enabling excellent inference performance even at high
accuracy. In MLPerf Inference v3.0, L4 delivered up to 3x more
performance than T4 at 99.9% of the reference (FP32) accuracy of
BERT—the highest BERT accuracy level tested in MLPerf Inference v3.0.
\n\n\n\nFigure 2. MLPerf Inference performance comparison between
NVIDIA L4 and NVIDIA T4 GPUs \n\nPer-accelerator throughput is not a
primary metric of MLPerf Inference. MLPerf Inference v3.0: Datacenter
Closed. Inference speedups calculated by dividing the inference
throughputs reported in MLPerf Inference v0.7 result ID 0.7-113 by
the number of accelerators to calculate T4 Tensor Core GPU per-
accelerator throughput and calculating the ratios of the inference
performance of the L4 Tensor Core GPU in \n\n
www.mlcommons.org\n\n3.0-0123 (Preview) by the calculated per-
accelerator throughput of T4. The MLPerf name and logo are trademarks
of MLCommons Association in the United States and other countries.
All rights reserved. Unauthorized use strictly prohibited. See for
more information. \n\nThe NVIDIA L4 also incorporates a large L2
cache, providing additional opportunity
to increase performance and energy efficiency. In the NVIDIA MLPerf
Inference v3.0 submission, two key software optimizations were
implemented to take advantage of the larger L2 cache: cache residency
and persistent cache management. \n\nThe larger L2 cache on L4
enabled the MLPerf workloads entirely within the cache. The L2 cache
provides higher bandwidth at lower power than GDDR memory, so the
significant reduction in GDDR accesses helped to both increase
performance and reduce energy use. \n\nUp to 1.4x higher performance
was observed when batch sizes were optimized to enable the workloads
to fit entirely within the L2 cache, compared to the performance when
batch sizes were set to maximum capacity. \n\nL2 cache
persistence\n\nNVIDIA Ampere architecture\n\nAnother optimization
used the feature first introduced in the . This enables developers,
with a single call to TensorRT, to tag a subset of the L2 cache so
that it can be prioritized for retention (that is, scheduled to be
evicted last). This feature is especially useful for inference when
working under a regime of residency, as developers can target the
memory being reused for layer activations across model execution,
dramatically reducing GDDR write bandwidth usage. \n\n'''
```

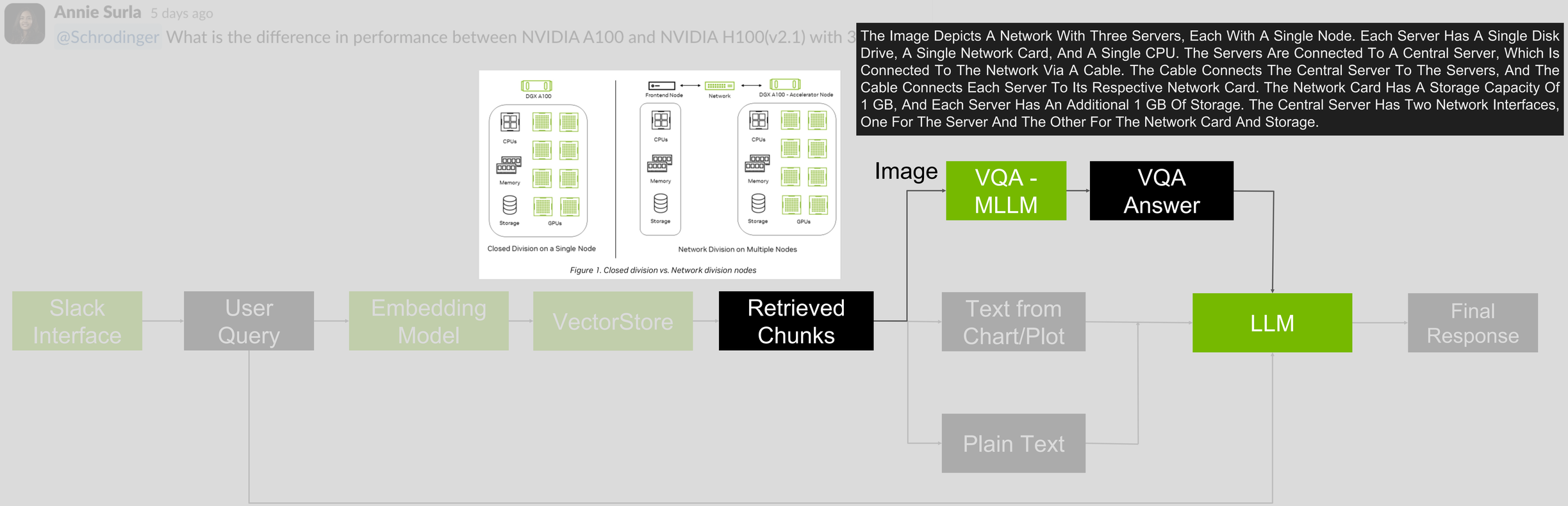
Multimodal RAG - Inference Workflow



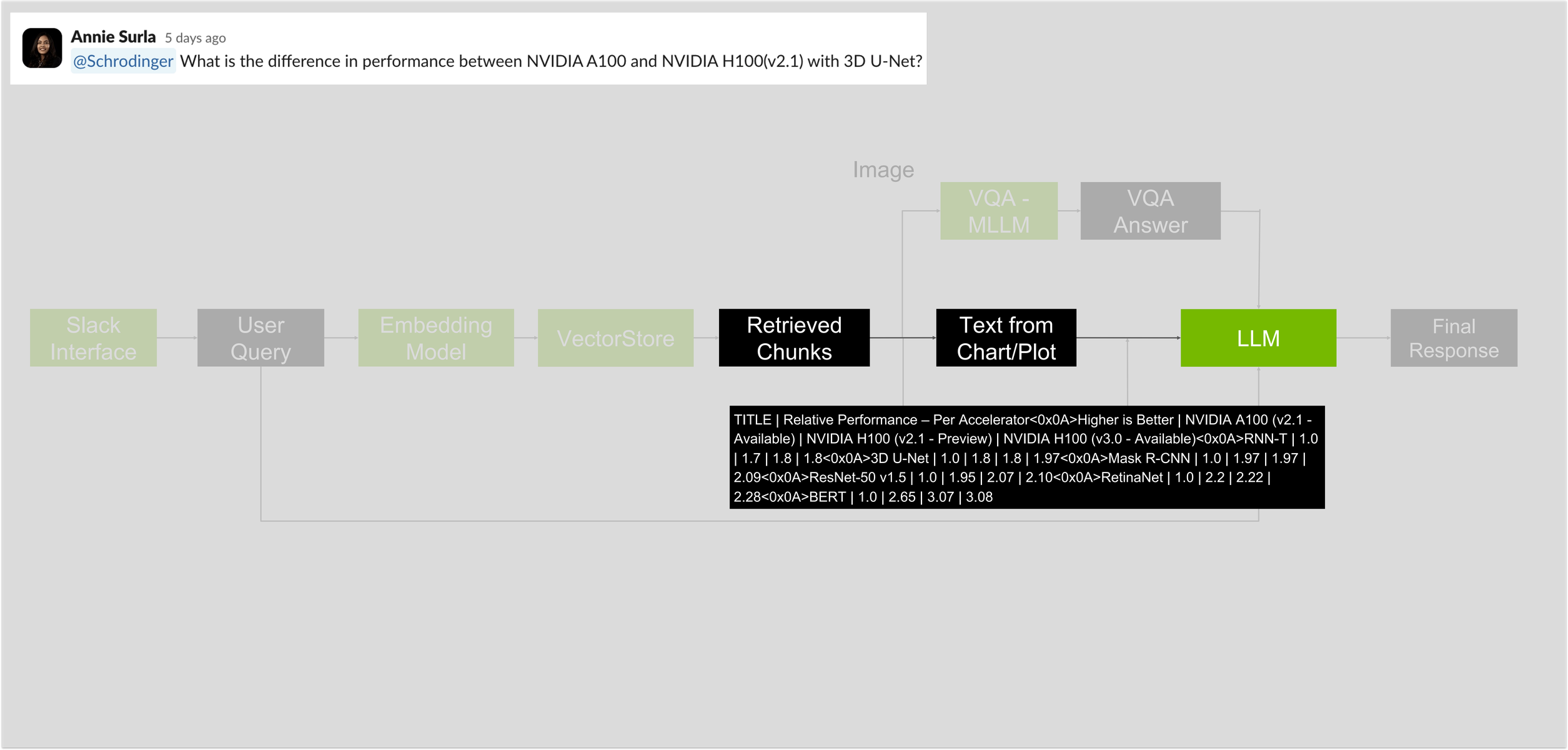
Multimodal RAG - Inference Workflow



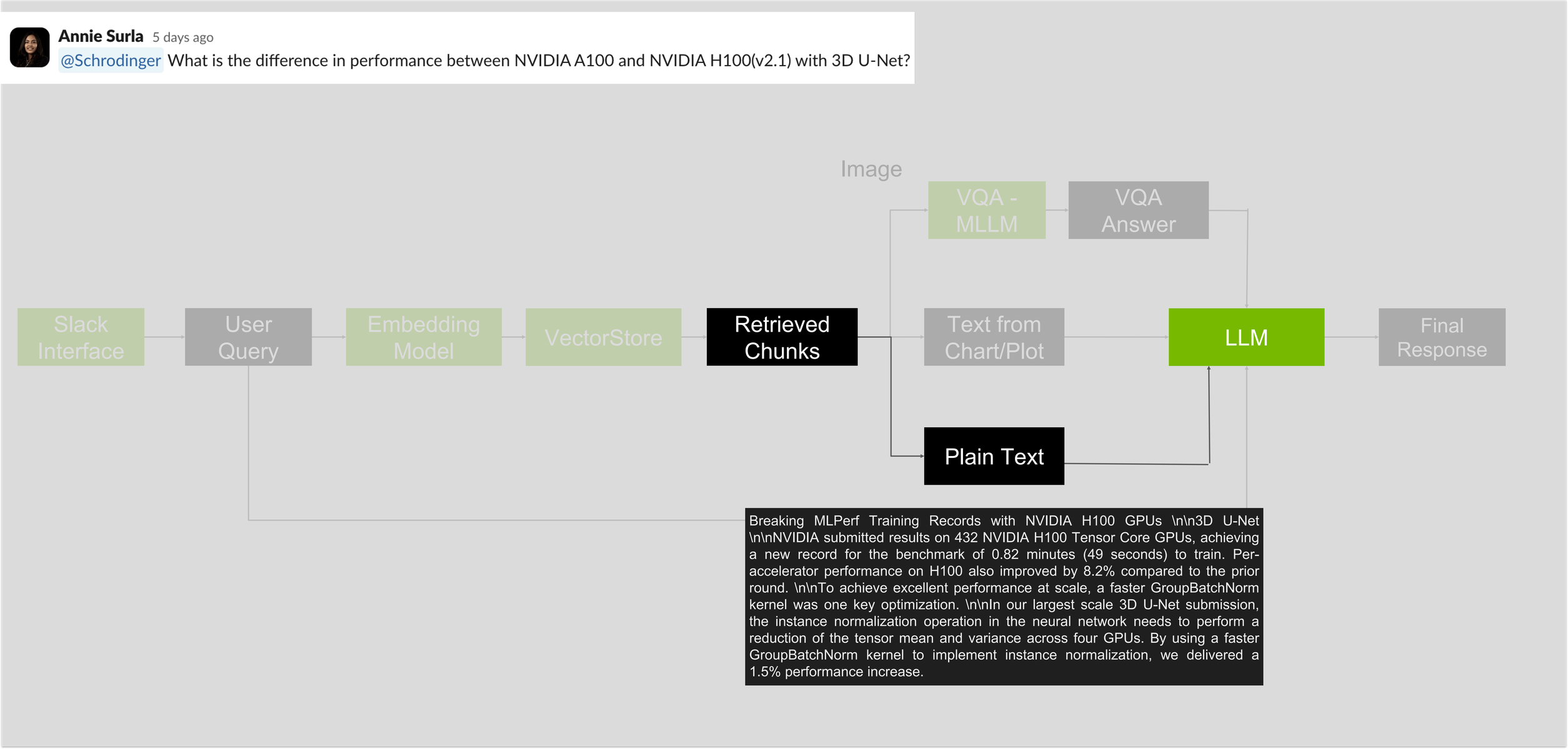
Multimodal RAG - Inference Workflow



Multimodal RAG - Inference Workflow



Multimodal RAG - Inference Workflow



Multimodal RAG - Inference Workflow

Annie Surla 5 days ago

@Schrodinger What is the difference in performance between NVIDIA A100 and NVIDIA H100(v2.1) with 3D U-Net?

6 replies

Schrodinger APP 5 days ago

The NVIDIA H100 (v2.1) has a relative performance of 1.8 with 3D U-Net, while the NVIDIA A100 has a performance of 1.0. This suggests that the NVIDIA H100 (v2.1) is 80% more efficient than the NVIDIA A100 when running 3D U-Net.

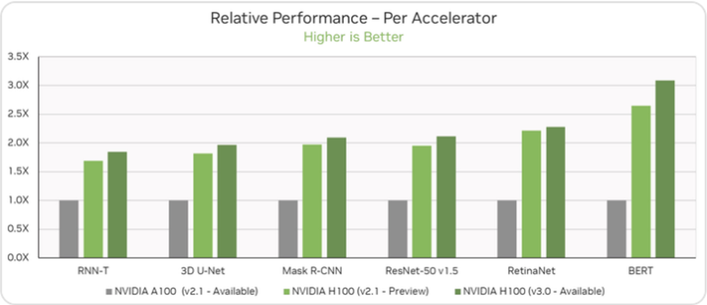
Sources Referred:

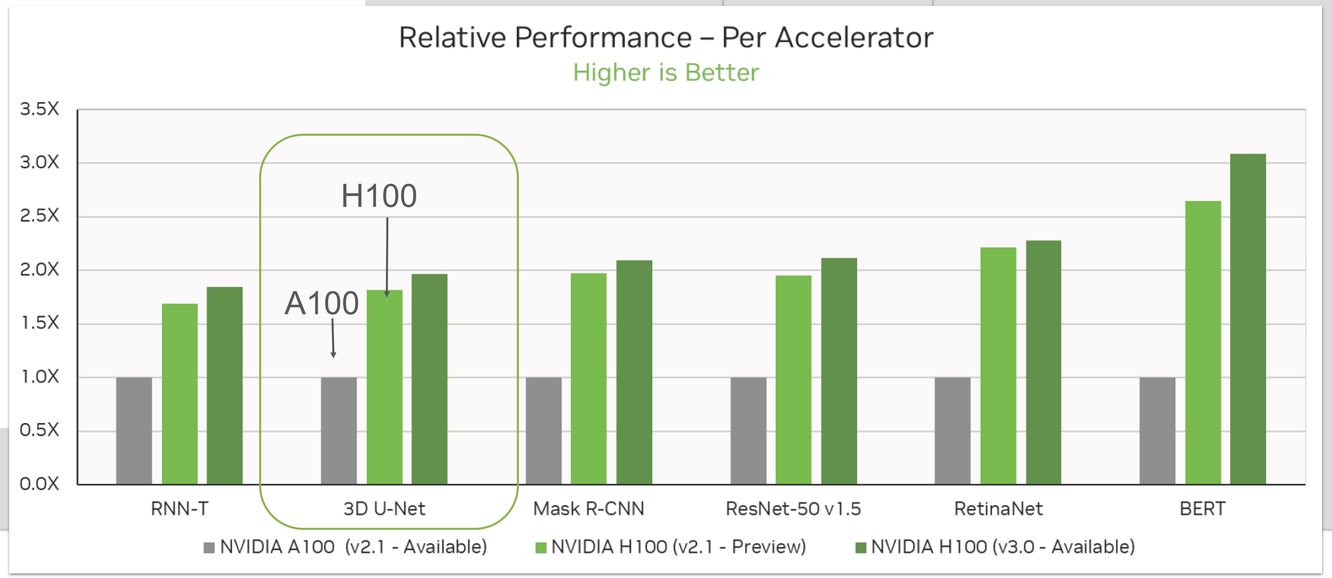
https://developer.nvidia.com/blog/breaking-mlperf-training-records-with-nvidia-h100-gpus/

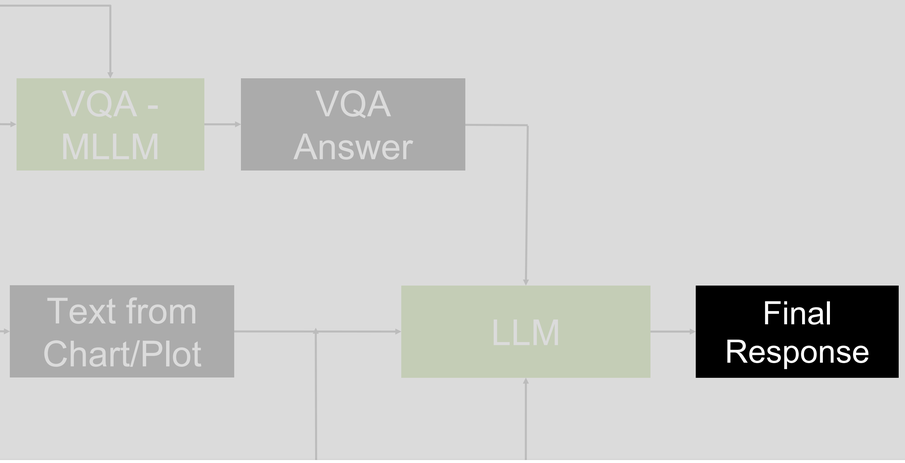
Schrodinger APP 5 days ago

Images referred from: https://developer.nvidia.com/blog/breaking-mlperf-training-records-with-nvidia-h100-gpus/

fileoutpart3

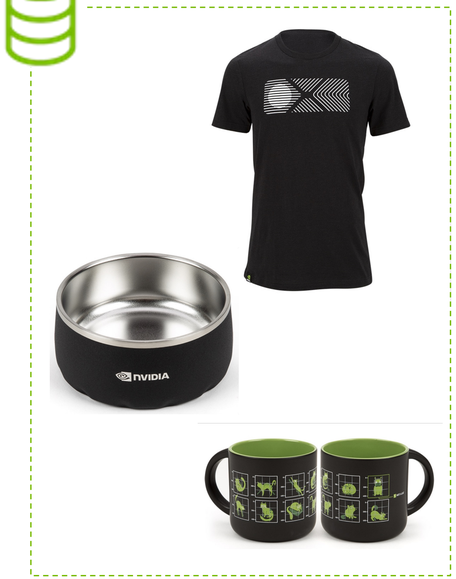




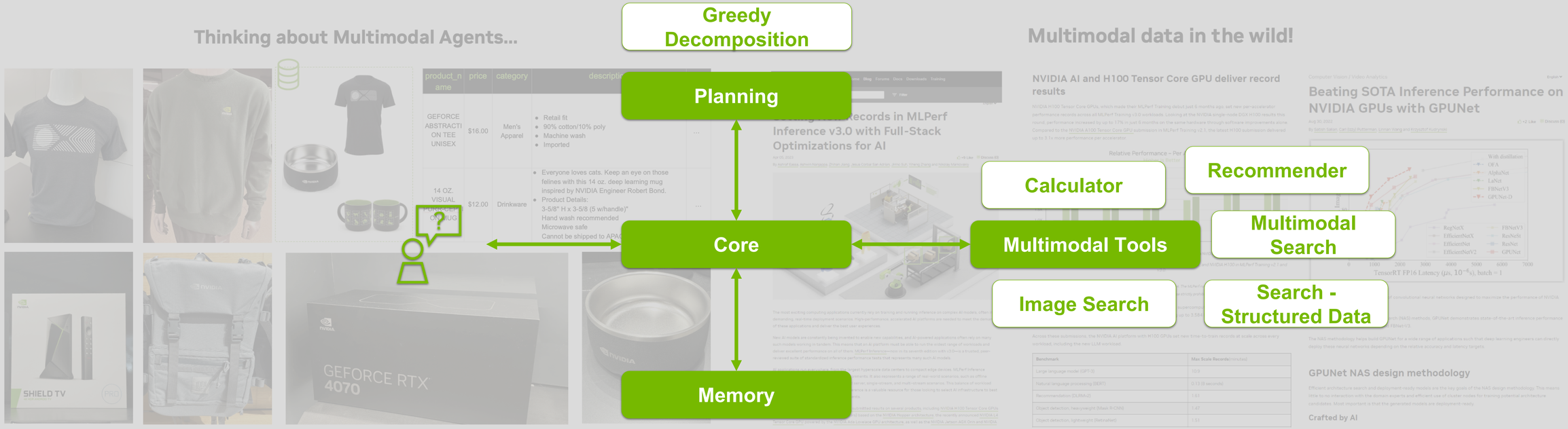


Thinking about Multimodal Agents...

product_name	price	category	description	…
GEFORCE ABSTRACTION TEE UNISEX	$16.00	Men's Apparel	• Retail fit • 90% cotton/10% poly • Machine wash • Imported	…
14 OZ. VISUAL PURR-CEPTION MUG	$12.00	Drinkware	• Everyone loves cats. Keep an eye on those felines with this 14 oz. deep learning mug inspired by NVIDIA Engineer Robert Bond. • Product Details: 3-5/8" H x 3-5/8 (5 w/handle)" Hand wash recommended Microwave safe Cannot be shipped to APAC	…



Building your first Multi-Modal Agent



Thinking about Multimodal Agents...

What material is this T-shirt made of?

User Question

Identify the Product Name in the User Query

Image Search

Generate an SQL Query based on Product Name

Search - Structured Data

Generate Final Answer

Response with Price

Search - Structured Data

Recommender

Tools

Image Search

Other API

Calculator/Math

Thinking about Multimodal Agents...

What material is this T-shirt made of?

User Question

Identify the Product Name in the User Query

Image Search

Generate an SQL Query based on Product Name

Search - Structured Data

Generate Final Answer

Response with Price

Search - Structured Data

Recommender

Tools

Image Search

Other API

Calculator/Math

Product Name User is referring to: GEFORCE ABSTRACTION TEE UNISEX

Thinking about Multimodal Agents...

What material is this T-shirt made of?

User Question

Identify the Product Name in the User Query

Image Search

Generate an SQL Query based on Product Name

Search - Structured Data

Generate Final Answer

Response with Price

Search - Structured Data

Recommender

Tools

Image Search

Other API

Calculator/Math

SELECT * FROM products WHERE product_name = "GEFORCE ABSTRACTION TEE UNISEX"
Retrieved Information: ('GEFORCE ABSTRACTION TEE UNISEX', '$16.00', "./Men's_files/NV00-0480-BKH_Full_t.jpg", 'https://gear.nvidia.com/GEFORCE-Abstraction-Tee-Unisex--P598.aspx', "Men's Apparel", 'Retail fit, **90% cotton, 10% poly**, Machine wash, Imported')

Thinking about Multimodal Agents...

What material is this T-shirt made of?

User Question

Identify the Product Name in the User Query

Image Search

Generate an SQL Query based on Product Name

Search - Structured Data

Generate Final Answer

Response with Price

Search - Structured Data

Recommender

Tools

Image Search

Other API

Calculator/Math

The GEFORCE ABSTRACTION T-SHIRT UNISEX is made of 90% cotton and 10% polyester.

What did we discuss today?

- RAG - is it enough?
- Agent - Answer complex business questions
  - A system with complex planning capabilities, memory, and the tools to execute tasks.
- Types of Agents
  - Data Agent
  - API Agent
- Multimodal RAG - How do we work with different modalities?
- Multimodal Agents - Chaining multimodal tools and working towards multimodal reasoning

Question?