基于NVIDIA AI FOUNDATION构建LLM-RAG有声检索智能体

NVIDIA企业级开发者社区 李奕澎

分享目录

- 开源的Mistral系列大模型
- NVIDIA Ai Foundation与NeMo微服务平台
- Langchain与NVIDIA Ai Foundation的结合使用
- 使用Llama Index构建RAG向量知识库
- 代码实战：构建语音交互的RAG智能体

MISTRAL大模型

Mistral 7B是一个拥有70亿参数的大模型，力压开源13B模型——Llama 2，并在推理能力、数学计算精准度以及代码生成任务上均超越了Llama 1 34B。其核心技术亮点包括采用分组查询注意力（Grouped-Query Attention, GQA）机制和滑动窗口注意力（Sliding Window Attention, SWA）策略，有效降低了推理阶段的资源消耗以显著提升推理速度。

Mistral AI_ Company

https://mistral.ai mistralai mistralai huggingface.co

AI & ML interests

None defined yet.

Team members 17

Models 5

- mistralai/Mixtral-8x7B-v0.1
  Text Generation • Updated 28 days ago • 193k • 1.3k
  → 专家混合模型版本，优于llama2 70B
- mistralai/Mistral-7B-Instruct-v0.2
  Text Generation • Updated Dec 15, 2023 • 624k • 889
  → Mistral-7B-Instruct-v0.1 的改进微调版本
- mistralai/Mixtral-8x7B-Instruct-v0.1
  Text Generation • Updated Dec 15, 2023 • 1.07M • 2.9k
  → 基于Mixtral 8x7B的微调版本
- mistralai/Mistral-7B-Instruct-v0.1
  Text Generation • Updated Dec 15, 2023 • 573k • 1.32k
  → Mistral 7B的微调版本
- mistralai/Mistral-7B-v0.1
  Text Generation • Updated Dec 12, 2023 • 952k • 2.79k
  → Mistral-Ai首版大模型，优于llama2 13B

GitHub - mistralai/mistral-src: Reference implementation of Mistral AI 7B v0.1 model.

MISTRAL模型特点

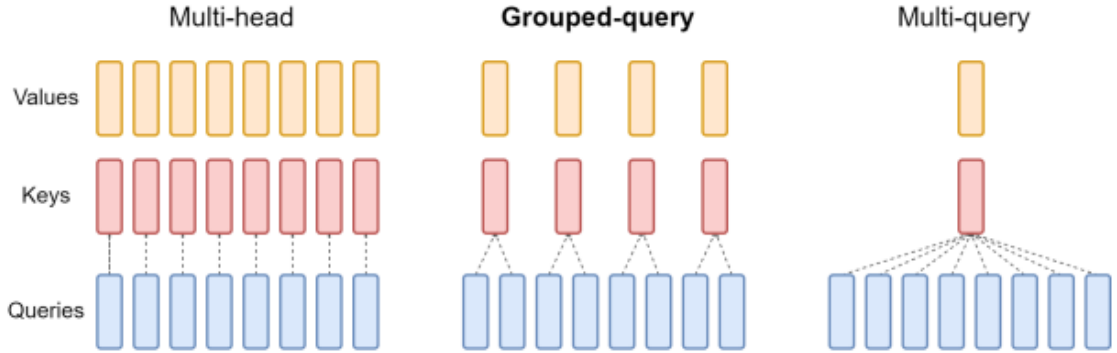
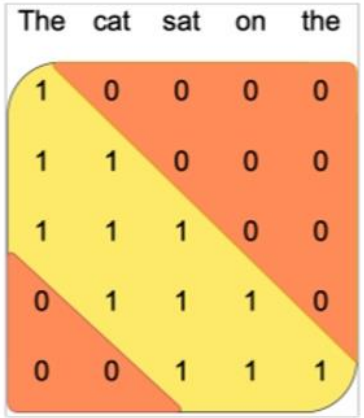


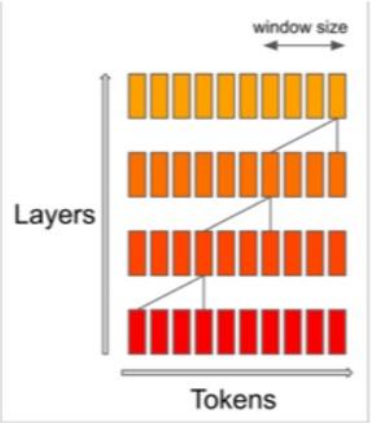
Figure 2: Overview of grouped-query method. Multi-head attention has H query, key, and value heads. Multi-query attention shares single key and value heads across all query heads. Grouped-query attention instead shares single key and value heads for each *group* of query heads, interpolating between multi-head and multi-query attention.

分组查询注意力 (Grouped-query attention)

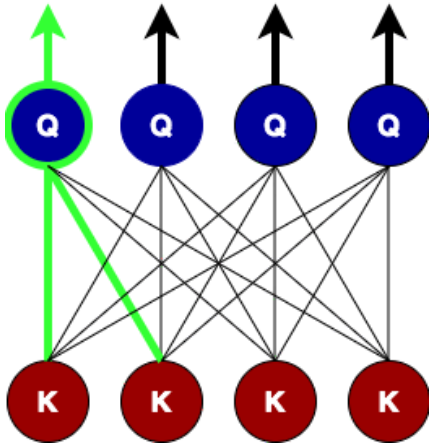
- 分组共享K,V参数减少缓存空间
- 加速推理



Sliding Window Attention



Effective Context Length



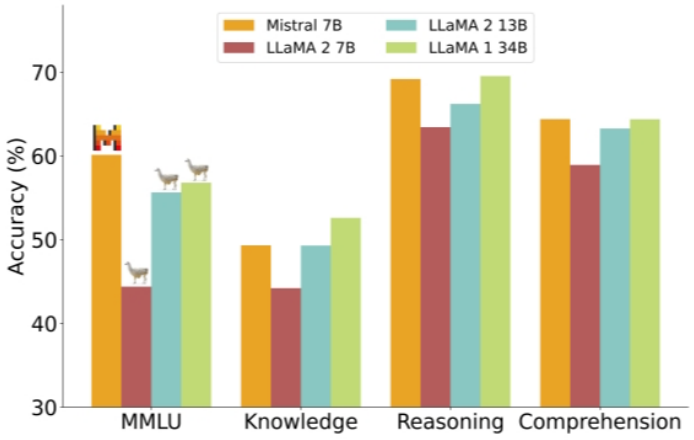
滑动窗口注意力 (Sliding Window Attention)

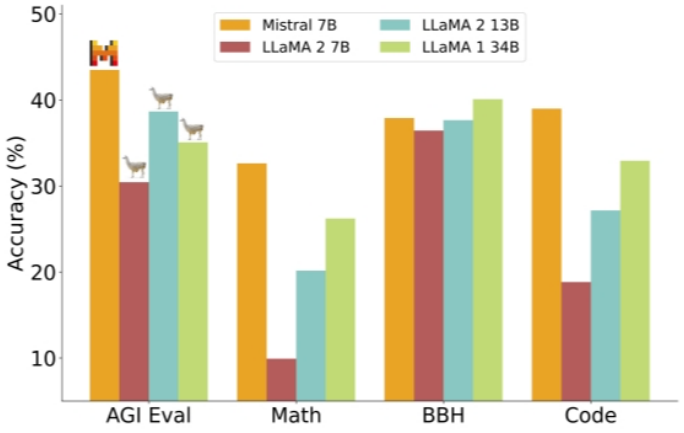
- 使用滑动窗口注意力进行长序列优化
- 实现更大的context上下文

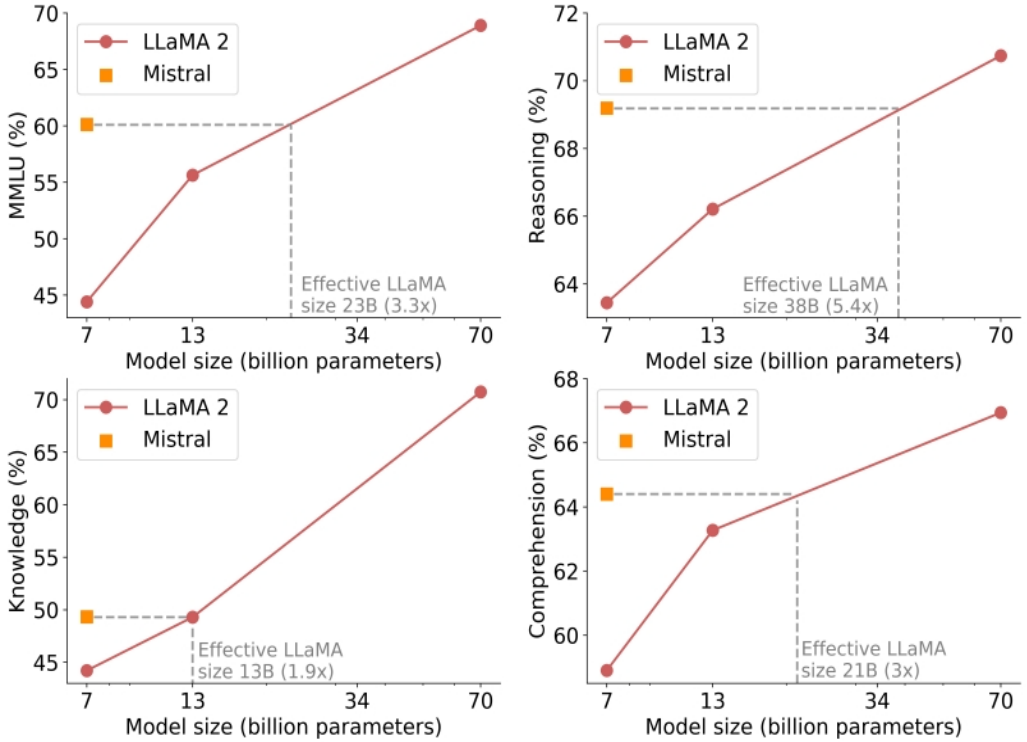
https://arxiv.org/pdf/2310.06825.pdf

GitHub - mistralai/mistral-src: Reference implementation of Mistral AI 7B v0.1 model.

MISTRAL大模型测评







Model	Modality	MMLU	HellaSwag	WinoG	PIQA	Arc-e	Arc-c	NQ	TriviaQA	HumanEval	MBPP	MATH	GSM8K
LLaMA 2 7B	Pretrained	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	24.7%	63.8%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	Pretrained	55.6%	**80.7%**	72.9%	80.8%	75.2%	48.8%	**29.0%**	**69.6%**	18.9%	35.4%	6.0%	34.3%
Code-Llama 7B	Finetuned	36.9%	62.9%	62.3%	72.8%	59.4%	34.5%	11.0%	34.9%	**31.1%**	**52.5%**	5.2%	20.8%
Mistral 7B	Pretrained	**60.1%**	**81.3%**	**75.3%**	**83.0%**	**80.0%**	**55.5%**	**28.8%**	**69.9%**	**30.5%**	47.5%	**13.1%**	**52.2%**

GitHub - mistralai/mistral-src: Reference implementation of Mistral AI 7B v0.1 model.

https://arxiv.org/pdf/2310.06825.pdf

NVIDIA AI Foundation

NVIDIA NGC — Ai大模型的创新乐园 GPU-optimized AI, Machine Learning, & HPC Software | NVIDIA NGC

NVIDIA. NGC | CATALOG

jetaimenoeda
ik8rqtofytxg

CATALOG
- Explore Catalog
- AI Foundation Models
- Collections
- Containers
- Helm Charts
- Models
- Resources

CONSOLE

ORGANIZATION

Collapse

NGC Catalog v1.139.0

Experience AI in Action

Experience state-of-the-art community and NVIDIA models optimized for peak performance and use them to kick-start your own development efforts.

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

NV-Llama2-70B-RLHF
Text Generation
NV-Llama2-70B-RLHF-Chat is a 70 billion parameter generative language model instruct-tuned on LLama2-70B model. It...
View Labels | Learn More

Maxine Live Portrait
2D Animation
Maxine Live Portrait is a generative model which animates a portrait photo with a driving video such that the facial...
View Labels | Learn More

Yi-34B
Text and Code Generation
The Yi-34B is a large language model trained from scratch by developers at 01.AI. Yi-34B has been finetuned for...
View Labels | Learn More

Nemotron-3-8B-Chat-SteerLM
Text Generation
Nemotron-3-8B-Chat-SteerLM is an 8 billion parameter generative language model based on the Nemotron-3-8B base...

NV-Llama2-70B-SteerLM-Chat
Text Generation
Llama 2 SteerLM Chat is a large language model, aligned using the SteerLM technique developed by NVIDIA. This...

Llama 2 70B
Text Generation
Llama 2 is a large language AI model capable of generating text and code in response to prompts.

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

NVIDIA AI Foundation

在 NVIDIA 加速基础设施上实现最佳性能，借助预训练的生成式人工智能模型，企业可以更快地创建自定义模型并利用最新的训练和推理技术。借助 NVIDIA AI Foundation Endpoints，应用程序可以连接到在完全加速的堆栈上运行的这些模型。

Language: Python | Response Type: No Streaming | Generate Key

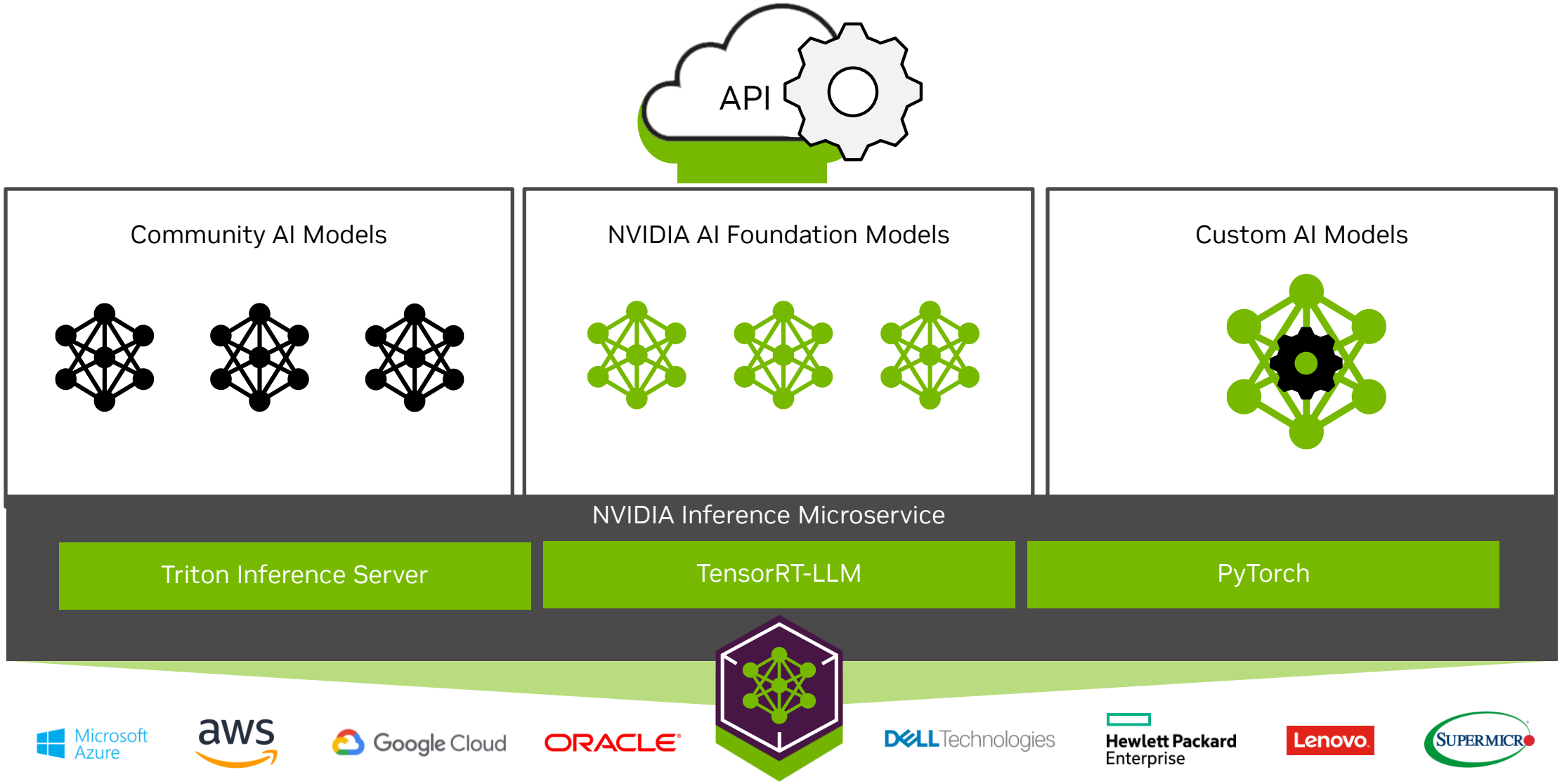
```python
import requests

invoke_url = "https://api.nvcf.nvidia.com/v2/nvcf/pexec/functions/35ec3354-2681-4d0e-a8dd-80325dcf7c63"
fetch_url_format = "https://api.nvcf.nvidia.com/v2/nvcf/pexec/status/"

headers = {
    "Authorization": "Bearer nvapi-MFjQKWXIwgAILtgErOG64UCUiR-zpa1UW_BiijLCDMMkML[redacted]",
    "Accept": "application/json",
}
```

```python
# re-use connections
session = requests.Session()

response = session.post(invoke_url, headers=headers, json=payload)

while response.status_code == 202:
    request_id = response.headers.get("NVCF-REQID")
    fetch_url = fetch_url_format + request_id
    response = session.get(fetch_url, headers=headers)

response.raise_for_status()
response_body = response.json()
print(response_body)
```

```python
payload = {
  "messages": [
    {
      "content": "I am going to Paris, what should I see?",
      "role": "user"
    }
  ],
  "temperature": 0.2,
  "top_p": 0.7,
  "max_tokens": 1024,
  "seed": 42,
  "stream": False
}
```

```
1. Eiffel Tower: This iconic landmark is a must-visit. You can take an elevat
or ride to the top for a stunning view of the city.
2. Louvre Museum: Home to thousands of works of art, including the Mona Lisa,
the Louvre is a must-visit for art lovers.
3. Notre-Dame Cathedral: This historic cathedral is a masterpiece of Gothic a
rchitecture and a must-see for anyone visiting Paris.
4. Montmartre: This artistic district is famous for its bohemian vibe, street
artists, and the beautiful Sacré-Cœur Basilica.
5. Champs-Élysées: This famous avenue is known for its luxury shops, cafes, a
nd the Arc de Triomphe.
6. Musée d'Orsay: This museum is home to an impressive collection of Impressi
onist and Post-Impressionist art.
7. Palace of Versailles: A short trip outside of Paris, the Palace of Versail
les is a must-visit for its opulent architecture and beautiful gardens.
```

NEMO MICROSERVICE



申请链接 Nemo Microservices EA Curator, Evaluator, and | NVIDIA Developer

LANGCHAIN-NVIDIA-AI-ENDPOINTS

With pip:

```
pip install langchain
```

With conda:

```
conda install langchain -c conda-forge
```

```
pip install -U langchain-nvidia-ai-endpoints
```

```
import getpass
import os

if not os.environ.get("NVIDIA_API_KEY", "").startswith("nvapi-"):
    nvapi_key = getpass.getpass("Enter your NVIDIA API key: ")
    assert nvapi_key.startswith("nvapi-"), f"{nvapi_key[:5]}... is not a valid key"
    os.environ["NVIDIA_API_KEY"] = nvapi_key
```

```
## Core LC Chat Interface
from langchain_nvidia_ai_endpoints import ChatNVIDIA

llm = ChatNVIDIA(model="mixtral_8x7b")
result = llm.invoke("Write a ballad about LangChain.")
print(result.content)
```


```
{'playground_nvolveqa_40k': '091a03bb-7364-4087-8090-bd71e9277520',
 'playground_nemotron_qa_8b': '0c60f14d-46cb-465e-b994-227e1c3d5047',
 'playground_mistral_7b': '35ec3354-2681-4d0e-a8dd-80325dcf7c63',
 'playground_seamless': '72ad9555-2e3d-4e73-9050-a37129064743',
 'playground_sdxl_turbo': '0ba5e4c7-4540-4a02-b43a-43980067f4af',
 'playground_sdxl': '89848fb8-549f-41bb-88cb-95d6597044a4',
 'playground_clip': '8c21289c-0b18-446d-8838-011b7249c513',
 'playground_yi_34b': '347fa3f3-d675-432c-b844-669ef8ee53df',
 'playground_llama_guard': 'b34280ac-24e4-4081-bfaa-501e9ee16b6f',
 'playground_deplot': '3bc390c7-eeec-40f7-a64d-0c6a719985f7',
 'playground_llama2_70b': '0e349b44-440a-44e1-93e9-abe8dcb27158',
 'playground_kosmos_2': '0bcd1a8c-451f-4b12-b7f0-64b4781190d1',
 'playground_fuyu_8b': '9f757064-657f-4c85-abd7-37a7a9b6ee11',
 'playground_nemotron_steerlm_8b': '1423ff2f-d1c7-4061-82a7-9e8c67afd43a',
 'playground_sd_video': 'a529a395-a7a0-4708-b4df-eb5e41d5ff60',
 'playground_llama2_code_70b': '2ae529dc-f728-4a46-9b8d-2697213666d8',
 'playground_neva_22b': '8bf70738-59b9-4e5f-bc87-7ab4203be7a0',
 'playground_cuopt': '8f2fbd00-2633-41ce-ab4e-e5736d74bff7',
 'playground_mixtral_8x7b': '8f4118ba-60a8-4e6b-8574-e38a4067a4a3',
 'playground_nv_llama2_rlhf_70b': '7b3e3361-4266-41c8-b312-f5e33c81fc92',
 'playground_llama2_code_34b': 'df2bee43-fb69-42b9-9ee5-f4eabbeaf3a8',
 'playground_llama2_code_13b': 'f6a96af4-8bf9-4294-96d6-d71aa787612e',
 'playground_llama2_13b': 'e0bb7fb9-5333-4a27-8534-c6288f921d3f',
 'playground_steerlm_llama_70b': 'd6fe6881-973a-4279-a0f8-e1d486c9618d'}
```

LANGCHAIN-NVIDIA-AI-ENDPOINTS

RAG检索增强生成

```
from langchain_community.vectorstores import FAISS
from langchain_core.output_parsers import StrOutputParser
from langchain_core.prompts import ChatPromptTemplate
from langchain_core.runnables import RunnablePassthrough
from langchain_nvidia_ai_endpoints import ChatNVIDIA
```

```
vectorstore = FAISS.from_texts(
    ["harrison worked at kensho"],
    embedding=NVIDIAEmbeddings(model="nvolveqa_40k"),
)
retriever = vectorstore.as_retriever()

prompt = ChatPromptTemplate.from_messages(
    [
        (
            "system",
            "Answer solely based on the following context:\n<Documents>\n{context}\n</Documents>",
        ),
        ("user", "{question}"),
    ]
)

model = ChatNVIDIA(model="mixtral_8x7b")

chain = (
    {"context": retriever, "question": RunnablePassthrough()}
    | prompt
    | model
    | StrOutputParser()
)

chain.invoke("where did harrison work?")
```

看图说话

```
import IPython
import requests

image_url = "https://www.nvidia.com/content/dam/en-zz/Solutions/research/ai-p
image_content = requests.get(image_url).content

IPython.display.Image(image_content)
```

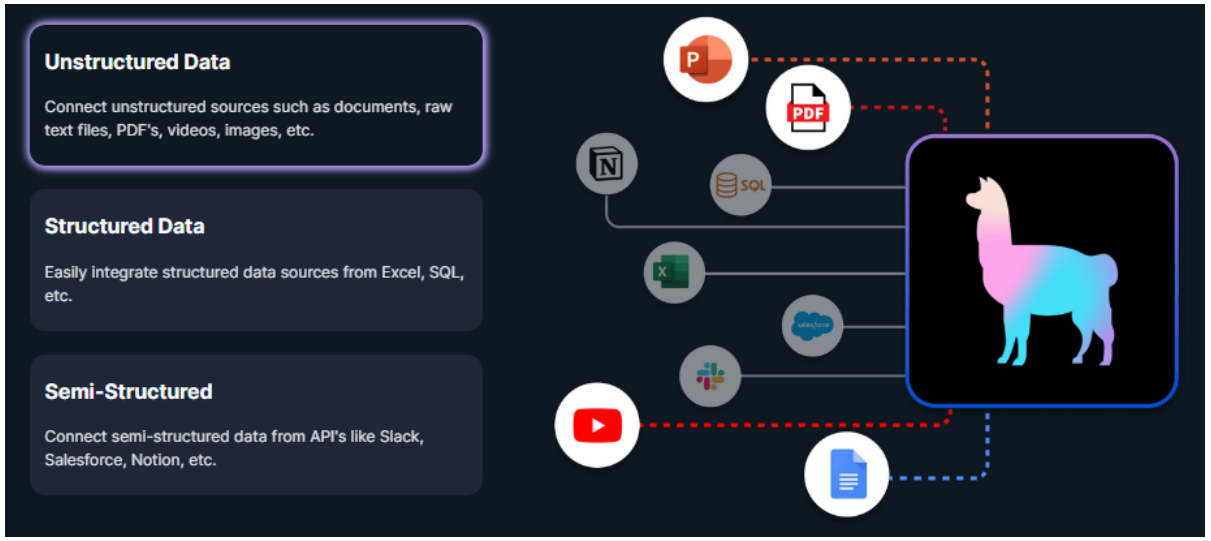
```
from langchain_nvidia_ai_endpoints import ChatNVIDIA

llm = ChatNVIDIA(model="playground_neva_22b")
```

```
from langchain_core.messages import HumanMessage

llm.invoke(
    [
        HumanMessage(
            content=[
                {"type": "text", "text": "Describe this image:"},
                {"type": "image_url", "image_url": {"url": image_url}},
            ]
        )
    ],
    labels={"creativity": 0, "quality": 9, "complexity": 0, "verbosity": 0},
)
```

```
Message(content='The image is a collage of three different pictures. The top picture features a cat with colorful,
```

Llama index工具库

LlamaIndex（以前称为 GPT Index）是一个数据框架，供 LLM 应用程序摄取、构建和访问私有或特定领域的数据。



数据读取
连接现有的数据源和数据格式（TXT、PDF、MD、SQL 等）以与大型语言模型应用程序一起使用。

数据索引
针对不同用例存储数据并为其建立索引。与下游矢量存储和数据库提供商集成。

查询接口
LlamaIndex 提供了一个查询接口，它接受数据上的任何输入提示并返回知识增强响应。

通过pip安装：

```
pip install llama-index
```

通过源码安装：

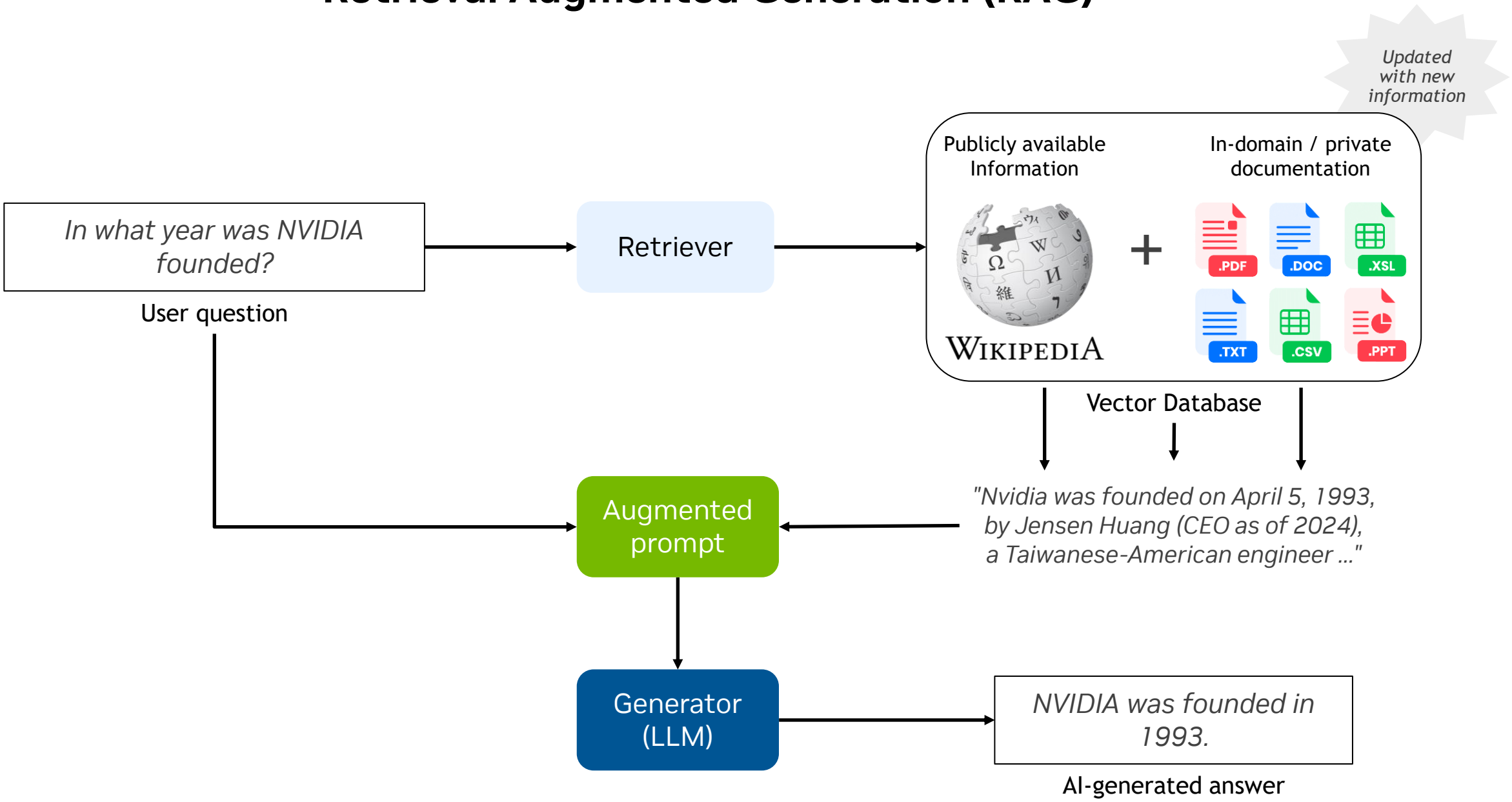
```
git clone https://github.com/jerryjliu/llama_index.git.
```

```
pip install -r requirements.txt
```

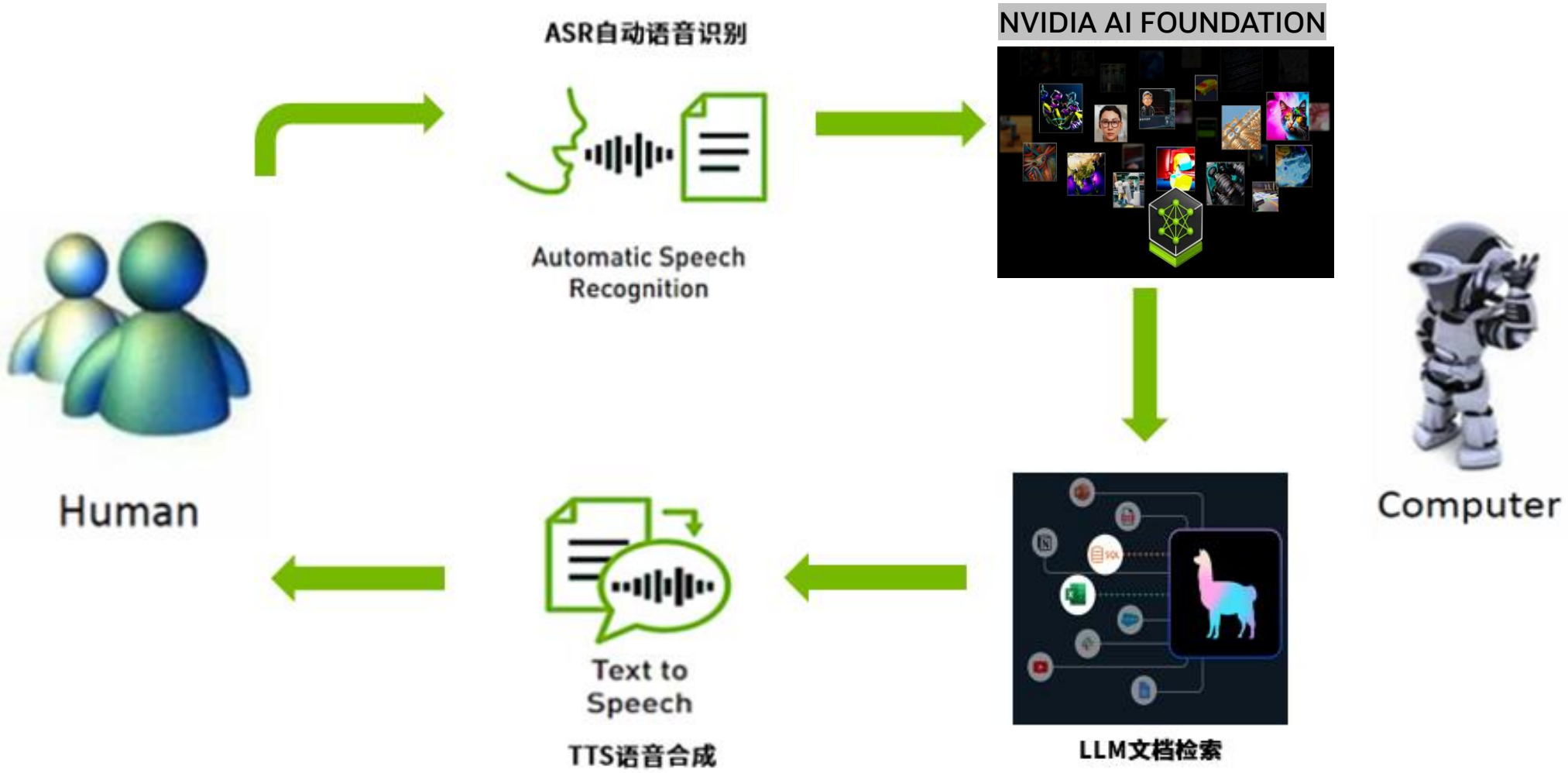
```
from llama_index import VectorStoreIndex, SimpleDirectoryReader

documents = SimpleDirectoryReader('data').load_data()
index = VectorStoreIndex.from_documents(documents)
```

```
query_engine = index.as_query_engine()
response = query_engine.query("What did the author do growing up?")
print(response)
```

Retrieval Augmented Generation (RAG)



LLM-RAG对话式AI交互



免费领取一门 NVIDIA DLI 自学课程

现在就加入 NVIDIA 开发者计划！

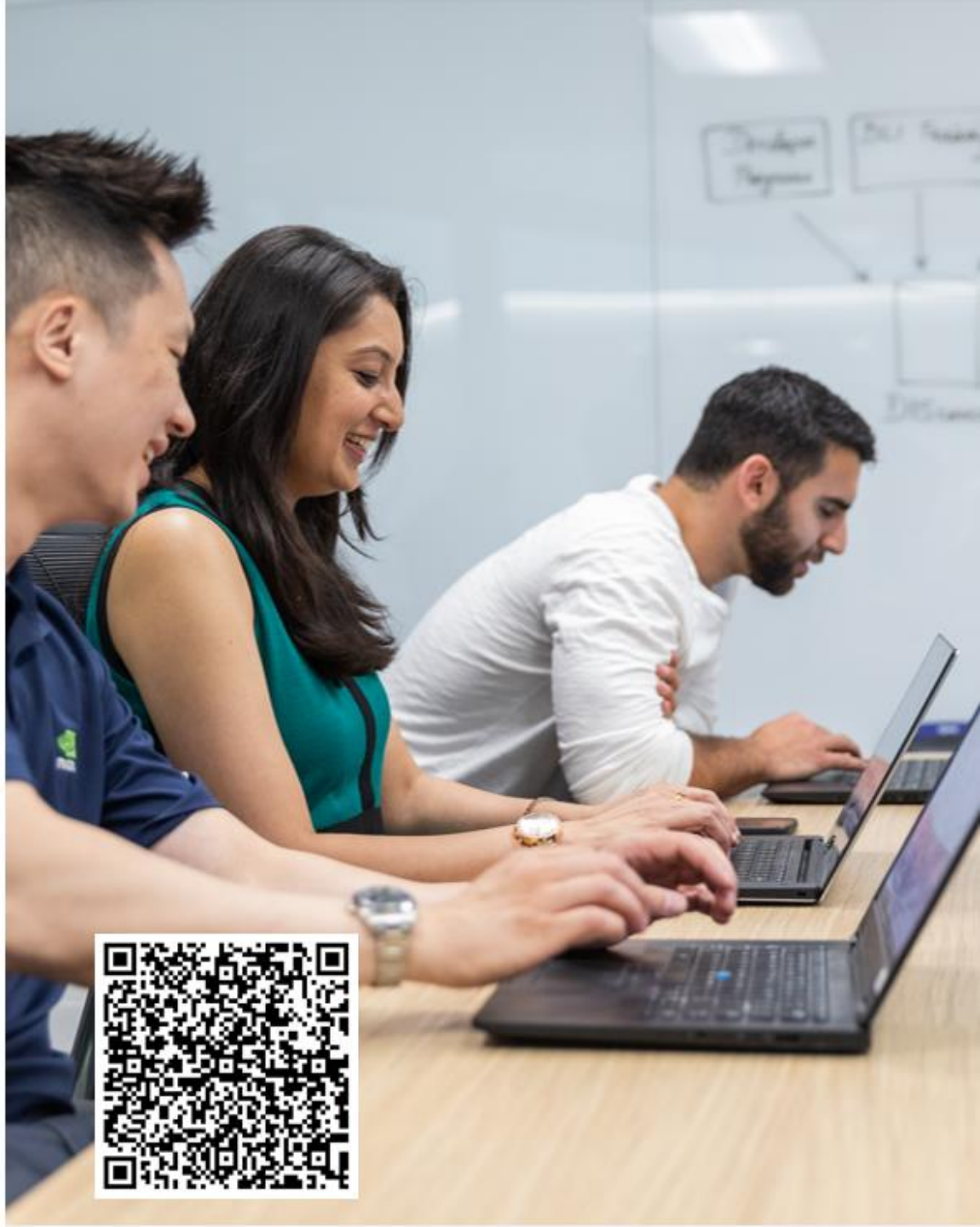
即日起，所有加入 **NVIDIA 开发者计划** 的新用户，即可免费领取一门 NVIDIA 深度学习培训中心 （DLI）自学课程。

NVIDIA 开发者计划权益：

- 通过 DLI 提供的技术培训和全球通行的培训证书**提升应用开发经验**。
- 免费访问 **150 多个SDKs** 和模型、在 GPU 上优化的软件、模型脚本和容器化的应用程序。
- 免费访问研究论文、技术文档、在线研讨会、开发者博客，以及 NVIDIA 最新资讯等**各类技术资源**。
- 在 **NVIDIA On-Demand** 上观看数千个技术会议视频。

微信扫码，领取福利！

（提醒：该福利仅面向新注册用户，已注册 NVIDIA 开发者计划的邮箱不能领取）



NVIDIA

THANK YOU

NVIDIA®