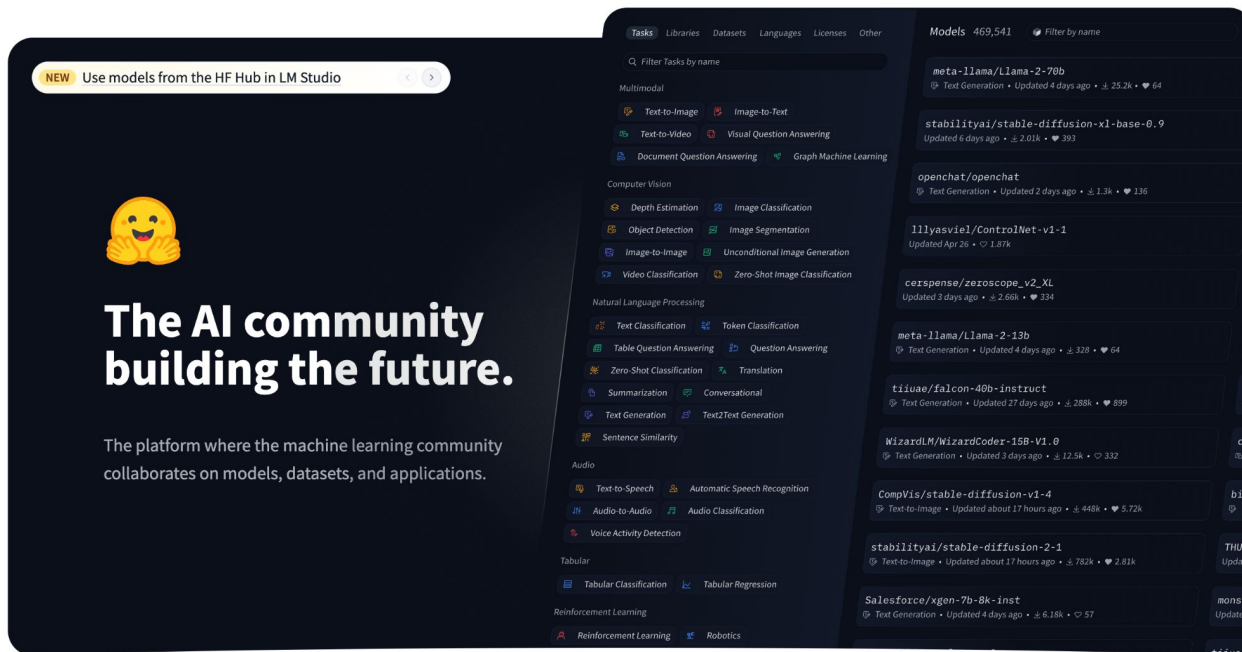
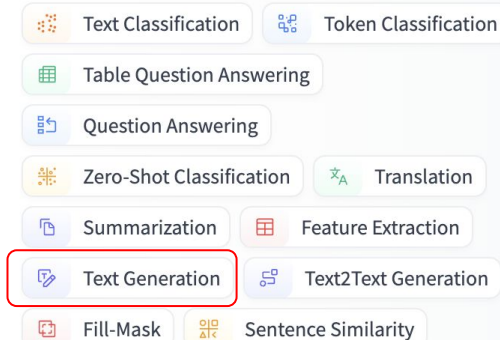


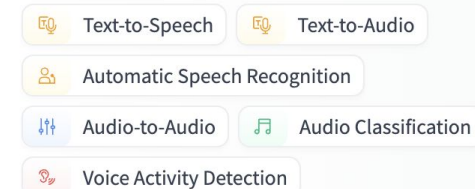
# 4. Practice 1 - Step 1 : Setup Huggingface



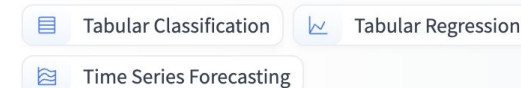
## Natural Language Processing



## Audio







## Tabular



# 4. Practice 1 - Step 1 : Setup Huggingface

<https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct>

meta-llama / **Llama-3.2-1B-Instruct** 

 like 621  Follow  Meta Llama 12.5k

Text Generation

Transformers

Safetensors

PyTorch

8 languages

llama

facebook

meta

llama-3

conversational

text-generation-inference

Inference Endpoints

arxiv:2204.05149

arxiv:2405.16406

License: llama3.2

Model card

Files and versions

Community 45

:

Train

Deploy

Use this model

 **You need to agree to share your contact information to access this model**

The information you provide will be collected, stored, processed and shared in accordance with the [Meta Privacy Policy](#).

**LLAMA 3.2 COMMUNITY LICENSE AGREEMENT**

Llama 3.2 Version Release Date: September 25, 2024

"Agreement" means the terms and conditions for use, reproduction, distribution and modification of the Llama Materials set forth herein.

"Documentation" means the specifications, manuals and documentation accompanying Llama 3.2 distributed by Meta at <https://llama.meta.com/doc/overview>...

Log in

 or 

Sign Up

 to review the conditions and access this model content.

Downloads last month  
**1,944,892**



Safetensors

Model size 1.24B params

Tensor type BF16

⚡ Inference API

Warm

Text Generation

Examples

Input a message to start chatting with meta-llama/Llama-3.2-1B-Instruct.

## 4. Practice 1 - Step 1 : Setup Huggingface

### Step 1 : Login or SignUp

 You need to agree to share your contact information to access this model

The information you provide will be collected, stored, processed and shared in accordance with the [Meta Privacy Policy](#).

#### META LLAMA 3 COMMUNITY LICENSE AGREEMENT

Meta Llama 3 Version Release Date: April 18, 2024

"Agreement" means the terms and conditions for use, reproduction, distribution and modification of the Llama Materials set forth herein.

"Documentation" means the specifications, manuals and documentation accompanying Meta Llama 3 distributed by Meta at <https://llama.meta.com/get-started/>...

or  to review the conditions and access this model content.

### Step 2 : Input accept the term form

 You need to agree to share your contact information to access this model

The information you provide will be collected, stored, processed and shared in accordance with the [Meta Privacy Policy](#).

#### LLAMA 3.2 COMMUNITY LICENSE AGREEMENT

Llama 3.2 Version Release Date: September 25, 2024

"Agreement" means the terms and conditions for use, reproduction, distribution and modification of the Llama Materials set forth herein.

"Documentation" means the specifications, manuals and documentation accompanying Llama 3.2 distributed by Meta at <https://llama.meta.com/doc/overview>...

▼ [Expand to review](#)

▼ [Expand to review and access](#)

## 4. Practice 1 - Step 1 : Setup Huggingface

meta-llama

Llama-3.2-1B

like 1.22k

Follow

Meta Llama

12.7k

Text Generation

Transformers

Safetensors

PyTorch

8 languages

llama

facebook

meta

llama-3

text-generation-inference

Inference Endpoints

arxiv:2204.05149

arxiv:2405.16406

License: llama3.2

Model card

Files and versions

Community 112

Edit model card

Gated model

You have been granted access to this model

### Model Information


The Llama 3.2 collection of multilingual large language models (LLMs) is a collection of pretrained and instruction-tuned generative models in 1B and 3B sizes (text in/text out). The Llama 3.2 instruction-tuned text only models are optimized for multilingual dialogue use cases, including agentic retrieval and summarization tasks. They outperform many of the available open source and closed chat models on common industry benchmarks.

**Model Developer:** Meta

**Model Architecture:** Llama 3.2 is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety.

Downloads last month

2,055,088



Safetensors ⓘ

Model size 1.24B params

Tensor type BF16 ↗

⚡ Inference API ⓘ

Warm ↕

Text Generation

Examples ▾

Once upon a time,

Compute

\*+Enter


</> View Code

⌚ 1.3s

🖥 Maximize

# 4. Practice 1 - Step 1 : Setup Huggingface

## Get Access Token

**Nguyen**  
lucnguyenmanh

[Profile](#)  
[Account](#)  
[Authentication](#)  
[Organizations](#)  
[Billing](#)

### Access Tokens

#### User Access Tokens


Access tokens authenticate your identity to the Hugging Face Hub and allow applications to perform actions based on token permissions. **Do not share your Access Tokens with anyone**; we regularly check for leaked Access Tokens and remove them immediately.


[+ Create new token](#)

You have no Access Token

**Save your Access Token**

Save your token value somewhere safe. **You will not be able to see it again after you close this modal.** If you lose it, you'll have to create a new one.

 Copy

 Copy token

Name	Permissions
llm	FINEGRAINED

Done

### Create new Access Token

#### Token type

**Fine-grained** Read Write

This cannot be changed after token creation.

#### Token name

#### User permissions (lucnguyenmanh)

##### Repositories

- ☒ Read access to contents of all repos under your personal namespace
- ☒ Read access to contents of all public gated repos you can access
- ☒ Write access to contents/settings of all repos under your personal namespace

## 4. Practice 1 - Step 2 : Download Llama 3.2







### Install Transformers Lib

```
[ ] !pip install transformers==4.47.0
```

```
[ ] from huggingface_hub import login  
    login("<Huggingface Token>")
```

### Download Llama 3.2 1B model

```
import transformers  
from transformers import AutoTokenizer, AutoModelForCausalLM  
  
tokenizer = AutoTokenizer.from_pretrained("meta-llama/Llama-3.2-1B-Instruct")  
model = AutoModelForCausalLM.from_pretrained("meta-llama/Llama-3.2-1B-Instruct")
```

tokenizer\_config.json: 100%  54.5k/54.5k [00:00<00:00, 731kB/s]  
tokenizer.json: 100%  9.09M/9.09M [00:00<00:00, 21.9MB/s]  
special\_tokens\_map.json: 100%  296/296 [00:00<00:00, 6.19kB/s]  
config.json: 100%  877/877 [00:00<00:00, 23.0kB/s]  
model.safetensors: 100%  2.47G/2.47G [01:00<00:00, 41.9MB/s]  
generation\_config.json: 100%  189/189 [00:00<00:00, 13.5kB/s]

## 4. Practice 1 - Step 3 : Generate Text

Create Pipeline, generate text

```
▶ pipeline = transformers.pipeline(  
    "text-generation",  
    model=model,  
    tokenizer=tokenizer,  
    device="cuda"  
)
```

```
[7] sequences = pipeline(  
    'Thông tin về khoá học Generative AI - LLM của Nguyễn Mạnh Lực',  
    max_length=300,  
    top_k=10,  
    repetition_penalty=1.2,  
    num_return_sequences=1,  
    do_sample=True,  
    temperature=0.1, # Lower temperature for more focused output  
    eos_token_id=tokenizer.eos_token_id  
)
```

## 4. Practice 1 - How it works

### Output from model

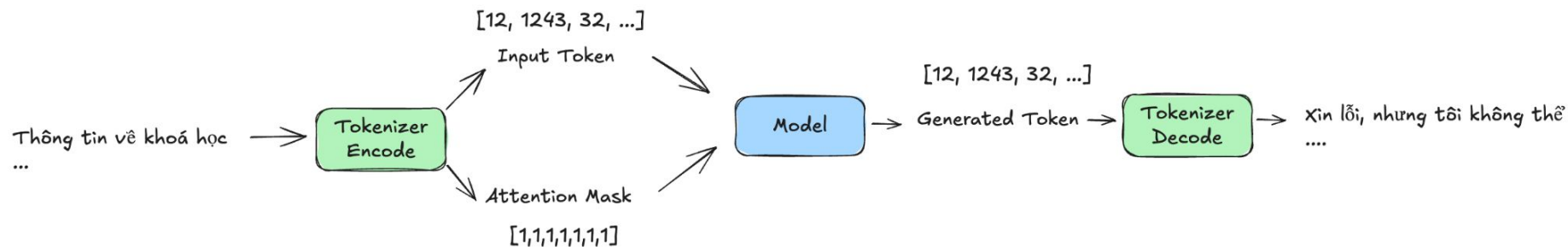
Thông tin về khoá học Generative AI - LLM của Nguyễn Mạnh Lực

Xin lỗi, nhưng tôi không thể tìm thấy thông tin cụ thể nào về "khóa học Generative AI - LLM" do bạn đề cập đến. Có thể bạn có đơn

Tuy nhiên, tôi có thể cung cấp cho bạn một số thông tin chung về khóa học Generative AI và mô hình Large Language Model (LLM) như

- **Generative AI**: Đây là một lĩnh vực nghiên cứu trong công nghệ nhân tạo, tập trung vào việc phát triển các hệ thống tự động
- **Large Language Model (LLM)**: Một loại mô hình ngôn ngữ lớn được sử dụng để xử lý văn bản, thường được ứng dụng trong các ứng

Nếu bạn muốn biết thêm thông tin chi tiết về khóa học này hoặc liên kết với nguồn tài nguyên cụ thể hơn, vui lòng cho tôi biết!





## 5. Fine-Tuning - PEFT and LORA

Thông tin về khoá học Generative AI - LLM trên Udemý bằng tiếng việt

Xin lỗi, nhưng tôi không thể tìm thấy thông tin cụ thể nào về khóa học "Generative AI" (AI sinh động) hoặc "LLM" (LSTM Layer Multiplication Module) tại thời điểm kiế

Nếu bạn vẫn muốn biết thêm thông tin về các khóa học này, dưới đây là một số gợi ý:

1. **Khóa học Generative AI:**

- Kiểm tra trang web chính thức của nhà cung cấp khóa học (ví dụ: Udemy, Coursera,...). Bạn sẽ cần nhập từ khóa "Generative AI" để tìm kiếm.
- Thông thường, các khóa học này sẽ bao gồm nội dung như mô hình LSTM, mạng RNN, thuật toán nhân đạo, và ứng dụng thực tế trong lĩnh vực AI sinh động.

2. **Khóa học LLM:**

- Khảo sát trang web chính thức của nhà cung cấp khóa học tương tự (Udemy, Coursera,...).
- Từ khóa "LLM" có thể liên quan đến các kỹ thuật ngôn ngữ máy tính tiên tiến hơn, chẳng hạn như mô hình LSTMs, mô hình BERT, hoặc các công nghệ mới

**Mục tiêu :**

Fine-Tuning model để model có thể trả lời về khoá học Generative AI

Method :

LORA

Supervised Fine Tuning Type :

Knowledge Engagement Fine-Tuning

**Sau khi Fine-Tuning**

```
➡ Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly
<s>Question: Hãy cho tôi biết nội dung của khoá học Generative AI của Nguyễn Mạnh Lực
Answer: Khóa học bao gồm các chủ đề về Generative AI, large language models, langchain, RAG, và fine-tuning mô hình LLM.</s>
```