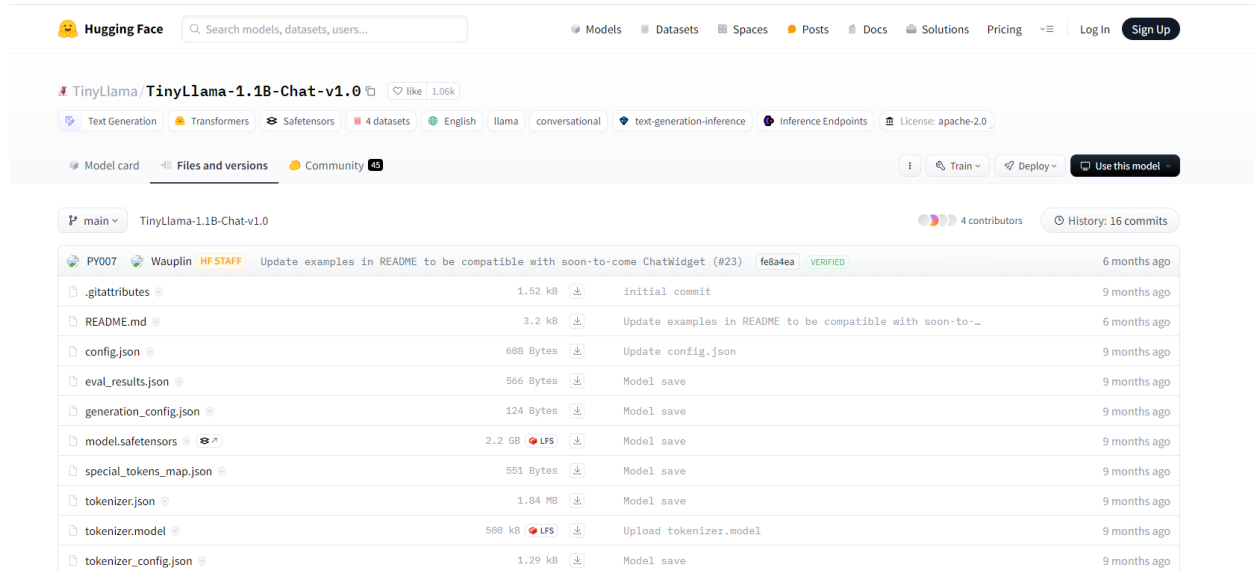


Instructions to running local LLMs

Go to <https://huggingface.co/> and find your model.



Hugging Face Search models, datasets, users...

Models Datasets Spaces Posts Docs Solutions Pricing Log In Sign Up

TinyLlama **TinyLlama-1.1B-Chat-v1.0** like 1.06k

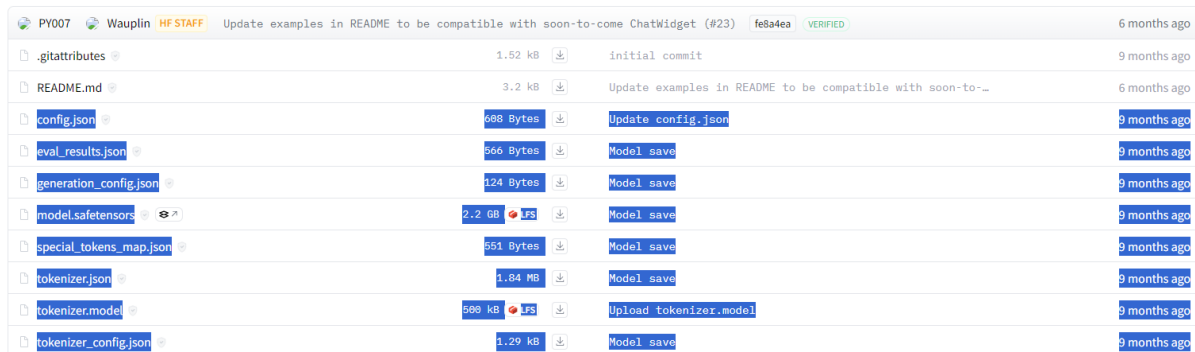
Text Generation Transformers Safetensors 4 datasets English llama conversational text-generation-inference Inference Endpoints License: apache-2.0

Model card Files and versions Community

main TinyLlama-1.1B-Chat-v1.0 4 contributors History: 16 commits

File	Size	Download	Commit	Time
.gitattributes	1.52 kB	Download	initial commit	9 months ago
README.md	3.2 kB	Download	Update examples in README to be compatible with soon-to-come ChatWidget (#23)	6 months ago
config.json	688 Bytes	Download	Update config.json	9 months ago
eval_results.json	566 Bytes	Download	Model save	9 months ago
generation_config.json	124 Bytes	Download	Model save	9 months ago
model.safetensors	2.2 GB	Download	Model save	9 months ago
special_tokens_map.json	551 Bytes	Download	Model save	9 months ago
tokenizer.json	1.84 MB	Download	Model save	9 months ago
tokenizer.model	508 kB	Download	Upload tokenizer.model	9 months ago
tokenizer_config.json	1.29 kB	Download	Model save	9 months ago

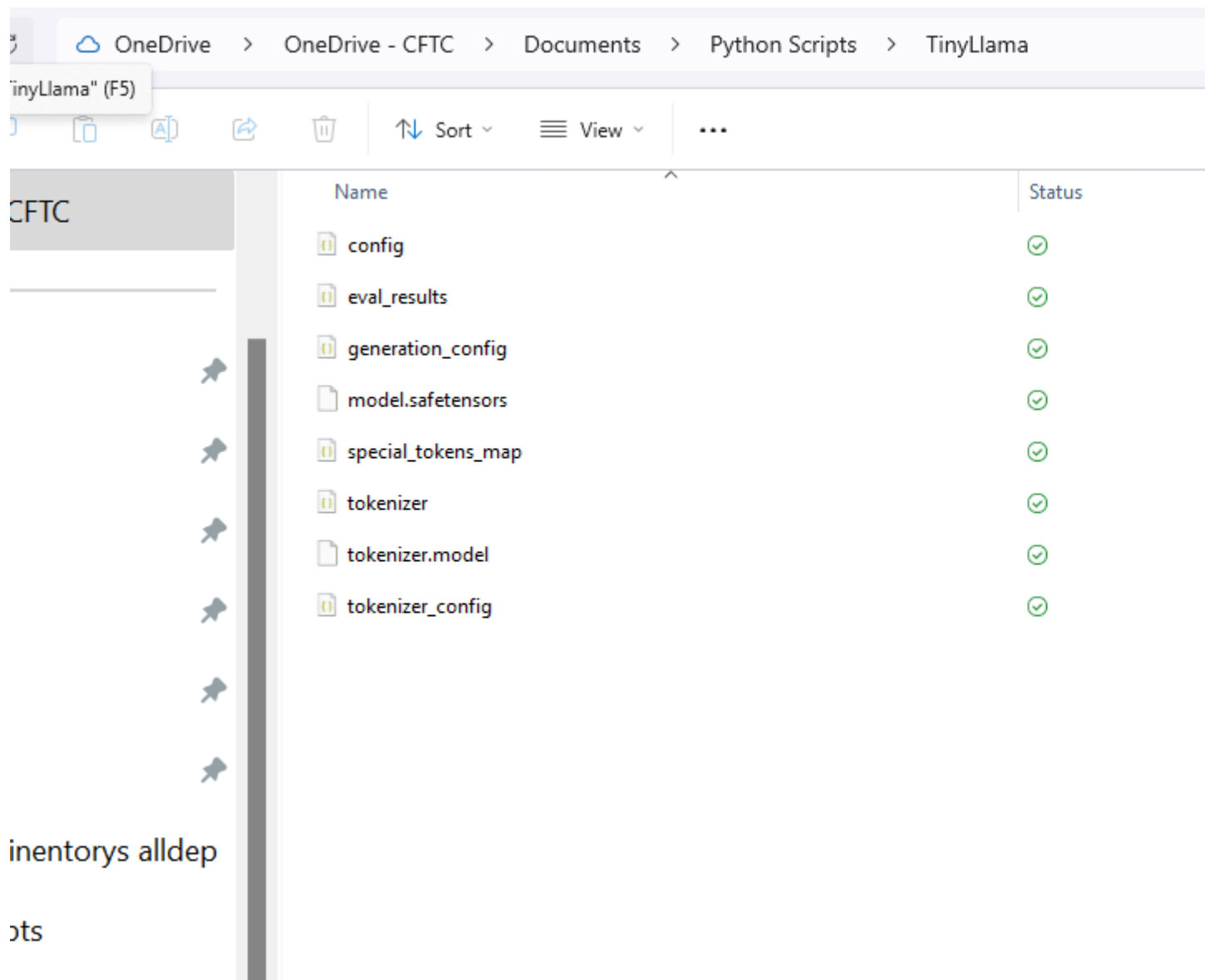
Download all necessary files.



PY007 **Wauplin** **HF STAFF** Update examples in README to be compatible with soon-to-come ChatWidget (#23) **fe8a4ea** **VERIFIED** 6 months ago

File	Size	Download	Commit	Time
.gitattributes	1.52 kB	Download	initial commit	9 months ago
README.md	3.2 kB	Download	Update examples in README to be compatible with soon-to-come ChatWidget (#23)	6 months ago
config.json	688 Bytes	Download	Update config.json	9 months ago
eval_results.json	566 Bytes	Download	Model save	9 months ago
generation_config.json	124 Bytes	Download	Model save	9 months ago
model.safetensors	2.2 GB	Download	Model save	9 months ago
special_tokens_map.json	551 Bytes	Download	Model save	9 months ago
tokenizer.json	1.84 MB	Download	Model save	9 months ago
tokenizer.model	508 kB	Download	Upload tokenizer.model	9 months ago
tokenizer_config.json	1.29 kB	Download	Model save	9 months ago

Choose a directory



Install packages

```
pip install torch transformers
```

May need to use conda-forge method bellow

```
conda install pytorch=1.9.0 torchvision torchaudio -c conda-forge
```

Load Packages

```
# Packages
import torch
from transformers import AutoTokenizer, AutoModelForCausalLM
```

Set the path

```
# Set the path to the directory containing the model files
model_path = "D:\Data\OneDrive\Ccantu\OneDrive - CFTC\Documents\Python Scripts\TinyLlama"
```

Load the model and tokenizer

```
# Load tokenizer and model
tokenizer = AutoTokenizer.from_pretrained(model_path)
model = AutoModelForCausalLM.from_pretrained(model_path, torch_dtype=torch.float16, device_map="auto")
```

Example generation function

```
def generate_response(prompt, max_length=100): # max_length keeps response short
    inputs = tokenizer(prompt, return_tensors="pt").to(model.device)
    outputs = model.generate(**inputs, max_length=max_length, num_return_sequences=1) # num_return_sequences gives only 1 output sequence
    return tokenizer.decode(outputs[0], skip_special_tokens=True) # Turns output into a readable text and skips padding to a length
```

Example prompt

```
# Example prompt case
prompt = "Write a short paragraph about commodities:"
response = generate_response(prompt)
print(f"Prompt: {prompt}\n")
print(f"Generated Response:\n{response}")
```