Modules ■ Model I/O ■ Language models ■ LLMs ■ Integrations ■ Hugging Face Local Pipelines

Hugging Face Local Pipelines

Hugging Face models can be run locally through the `HuggingFacePipeline` class.

The Hugging Face Model Hub hosts over 120k models, 20k datasets, and 50k demo apps (Spaces), all open source and publicly available, in an online platform where people can easily collaborate and build ML together.

These can be called from LangChain either through this local pipeline wrapper or by calling their hosted inference endpoints through the HuggingFaceHub class. For more information on the hosted pipelines, see the HuggingFaceHub notebook.

To use, you should have the `transformers` python package installed.

```
pip install transformers > /dev/null
```

Load the model

```
from langchain import HuggingFacePipeline

llm = HuggingFacePipeline.from_model_id(
    model_id="bigscience/bloom-1b7",
    task="text-generation",
    model_kwargs={"temperature": 0, "max_length": 64},
)
```

```
    WARNING:root:Failed to default session, using empty session: HTTPConnectionPool(host='localhost', port=8000): Max retries exceeded with url: /sessions (Caused by NewConnectionError('<urllib3.connection.HTTPConnection object at 0x1117f9790>: Failed to establish a new connection: [Errno 61] Connection refused'))
```

Integrate the model in an LLMChain

```python
from langchain import PromptTemplate, LLMChain

template = """Question: {question}

Answer: Let's think step by step."""
prompt = PromptTemplate(template=template, input_variables=["question"])

llm_chain = LLMChain(prompt=prompt, llm=llm)

question = "What is electroencephalography?"

print(llm_chain.run(question))
```

```
    /Users/wfh/code/lc/lckg/.venv/lib/python3.11/site-packages/transformers/generation/utils.py:1288: UserWarning: Using `max_length`'s default (64) to control the generation length. This behaviour is deprecated and will be removed from the config in v5 of Transformers -- we recommend using `max_new_tokens` to control the maximum length of the generation.
      warnings.warn(
    WARNING:root:Failed to persist run: HTTPConnectionPool(host='localhost', port=8000): Max retries exceeded with url: /chain-runs (Caused by NewConnectionError('<urllib3.connection.HTTPConnection object at 0x144d06910>: Failed to establish a new connection: [Errno 61] Connection refused'))
```

```
     First, we need to understand what is an electroencephalogram. An electroencephalogram is a recording of
brain activity. It is a recording of brain activity that is made by placing electrodes on the scalp. The
electrodes are placed
```