OpenLLM

OpenLLM helps developers run any open-source LLMs, such as Llama 2 and Mistral, as OpenAI-compatible API endpoints, locally and in the cloud, optimized for serving throughput and production deployment.

- GitHub: https://github.com/bentoml/OpenLLM
- CoLab: https://colab.research.google.com/github/bentoml/OpenLLM/blob/main/examples/llama2.ipynb

Install

> Recommend using a Python Virtual Environment

```
pip install openllm
```

Start a LLM Server

```
openllm start microsoft/Phi-3-mini-4k-instruct --trust-remote-code
```

To interact with the server, you can visit the web UI at http://localhost:3000/ or send a request using curl. You can also use OpenLLM's built-in Python client to interact with the server:

```
import openllm

client = openllm.HTTPClient('http://localhost:3000')
client.generate('Explain to me the difference between "further" and "farther"')
```

OpenAI Compatible Endpoints

```
import openai

client = openai.OpenAI(base_url='http://localhost:3000/v1', api_key='na')  # Here the server is running on 0.0.0.0:3000
```

```
completions = client.chat.completions.create(
  prompt='Write me a tag line for an ice cream shop.', model=model, max_tokens=64,
stream=stream
)
```

LangChain

```
from langchain.llms import OpenLLMAPI

llm = OpenLLMAPI(server_url='http://44.23.123.1:3000')
llm.invoke('What is the difference between a duck and a goose? And why there are so many Goose
in Canada?')

# streaming
for it in llm.stream('What is the difference between a duck and a goose? And why there are so
many Goose in Canada?'):
  print(it, flush=True, end='')

# async context
await llm.ainvoke('What is the difference between a duck and a goose? And why there are so
many Goose in Canada?')

# async streaming
async for it in llm.astream('What is the difference between a duck and a goose? And why there
are so many Goose in Canada?'):
  print(it, flush=True, end='')
```

Revision #3
Created 2024-06-21 11:00:23 CST by A-Lang (Admin)
Updated 2024-06-21 11:16:01 CST by A-Lang (Admin)