Testing AUP Remote Learning Labs LLM Serving Infrastructure

This document describes how to test the LLM serving infrastructure hosted in the AUP Remote learning labs.

This document assumes that your IP address has already been enabled in the system.

Table of Content

Ollama Endpoint

You can check that Ollama is running by accessing this URL in a web browser: https://aupcloud.io/ollama-endpoint/ and you should see the message `Ollama is running`.

CLI

Run inference from the CLI

```
curl -X POST https://aupcloud.io/ollama-endpoint/api/generate -d '{"model":
"llama3.1:8b", "prompt": "Why is the sky blue?", "stream": false}'
```

Ollama Python API

Run inference using the Ollama Python API

First, make sure the Ollama package is installed

```
python3 -m pip install ollama
```

List of Models

```
import ollama

client = ollama.Client(host='https://aupcloud.io/ollama-endpoint')
models_info = client.list()
model_list = [model.model for model in models_info['models']]

print(model_list)
```

Run Inference

```
from ollama import Client

client = Client(
    host='https://aupcloud.io/ollama-endpoint'
)
response = client.chat(model='llama3.1:8b', messages=[
  {
    'role': 'user',
    'content': 'Why is the sky blue?',
  },
])

print(f'{response.message.content=}')
```

OpenAI Python API

Run inference using the OpenAI Python API

First, make sure the OpenAI package is installed

```
python3 -m pip install openai
```

```
from openai import OpenAI

client = OpenAI(
    base_url='https://aupcloud.io/ollama-endpoint/v1',
    api_key='ollama',
)

response = client.chat.completions.create(
  model="llama3.1:8b",
  messages=[
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": "Who won the world series in 2020?"},
    {"role": "assistant", "content": "The LA Dodgers won in 2020."},
    {"role": "user", "content": "Where was it played?"}
  ]
)

print(response.choices[0].message.content)
```

LibreChat