Клиент для локальной модели с тулом

Обработка естественного языка
Практика 6

Михаил Пожидаев

Институт прикладной математики и компьютерных наук ТГУ

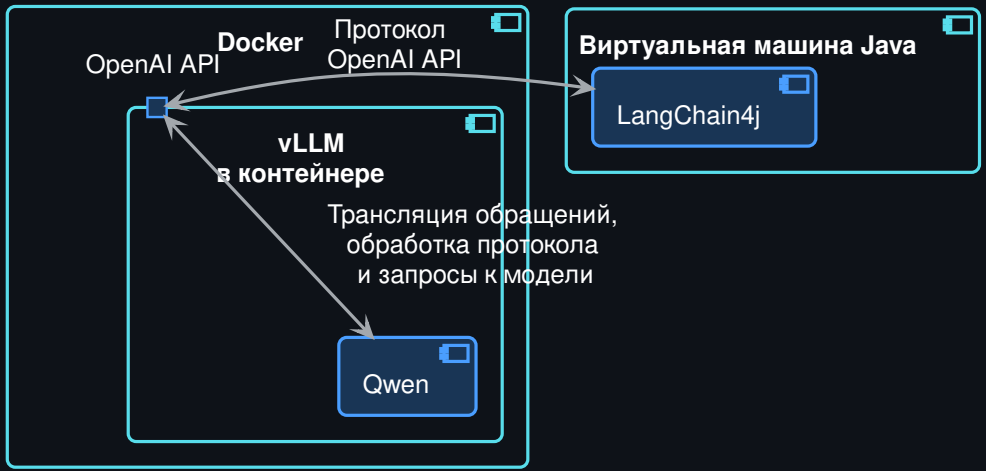
OpenAI API

Унифицированный JSON-like интерфейс к языковым моделям

OpenAI API** позволяет интегрировать возможности языковых моделей в сайты, приложения и внутренние сервисы. С его помощью можно создавать чат-ботов, генерировать тексты, анализировать документы и пр

Полезные свойства

- поддерживает вызов внешних функций (tool calling).
- Имеет множество клиентов на разных языках и на разные случаи жизни
- Изолирует логику клиентов от особенностей поведения разных языковых моделей.
- Поддерживает стриминговый режим работы.


Docker
OpenAI API
Протокол
OpenAI API
Виртуальная машина Java
LangChain4j
vLLM
в контейнере
Трансляция обращений,
обработка протокола
и запросы к модели
Qwen

Запуск VLLM

Команда Docker для запуска VLLM в CPU-режиме

```
docker run --rm --name vllm -d \
  --ipc=host \
  -p 8000:8000 \
  -v /tmp/huggingface:/root/.cache/huggingface \
  vllm/vllm-openai-cpu:latest \
  --model Qwen/Qwen2.5-7B-Instruct \
    --gpu-memory-utilization 0.50 --max-model-len 4096 \
    --enable-auto-tool-choice --tool-call-parser hermes
```

Проверка VLLM

Обработка пробного запроса с использованием `curl`

```
curl -s http://localhost:8000/v1/chat/completions \
  -H "Content-Type: application/json" -H "Authorization: Bearer dummy" \
  -d '{
    "model": "Qwen/Qwen2.5-7B-Instruct",
    "messages": [
      {
        "role": "user", "content": "Write a funny story about stupid language models"
      }]}' | jq .
```

Спасибо за внимание!

Домашняя страница:
https://marigostra.ru

Электронная почта:
msp@luwrain.org

Telegram:
@Marigostra