framework to set up the server-side. On the client-side, data is transmitted using the request-response structure to achieve API access. In both the server and client implementations of RKLLM-Server-Flask, special consideration is given to the calling method of the OpenAI-API. The example code ensures consistency by setting up data structures identical to those used in the OpenAI-API. This allows users to seamlessly migrate their services by simply replacing the access interface after deploying RKLLM-Server-Flask, leveraging their existing development foundations in OpenAI-API development.

According to the OpenAI-API usage documentation, users send specific data structures to the server during the API call process. The main contents are as follows:

```
{
    "model": "No models available",
    "messages": [
      {"role": "system",
        "content": "You are a helpful assistant."
      },
      {
        "role": "user",
        "content": "Hello!"
      }
    ],
    "stream": false,
}
```

Among them, "model" and "stream" specify the specific model to be called and whether to initiate streaming inference transmission, while the "content" data in "messages" is the crucial user input.

As for the data returned by the server, the structure of the data output by the OpenAI-API varies depending on whether streaming inference transmission is selected. The data content returned under non-streaming inference settings is as follows: