# /api/chat API文档

本API采用HTTP POST方法进行请求，并需要设置Authorization HTTP header为token。

## 请求格式

请求格式如下：

POST /api/chat HTTP/1.1  
Authorization: Bearer YOUR\_TOKEN  
  
{  
 "model": "openbuddy-llama-30b-v7.1-bf16",  
 "messages": [  
 {  
 "role": "user",  
 "content": "test"  
 }  
 ],  
 "temperature": 0.5,  
 "max\_new\_tokens": 300,  
 "conversation\_id": getConversationID(),  
 "user\_id": getUserID(),  
 "system": "You are a helpful AI assistant......"  
}

## 请求参数

请求参数包含以下字段：

* model: 选择的模型名称。
* messages: 会话消息数组，每个元素是一个对象，包含角色（role，可以是“user”或“assistant”）和内容（content）。
* temperature: 控制AI生成结果的随机性的值，范围在0和0.9之间，值越大结果越具备随机性、越有创意。
* max\_new\_tokens: AI在每个响应中最多生成的新token数。
* conversation\_id: 对话ID，由客户端生成。必须是UUID的格式。
* user\_id: 应用的用户ID，用于审查。
* system: System Prompt，用于设定AI的角色和行为规范。

在标准的问答场景中，messages数组的最后一项的role应当为user，此时，模型会以assistant的身份输出新消息回答用户。

在需要续写AI的回答的场景中，最后一项的role应当为assistant，之后模型会在之前的回答的基础上续写，并只输出续写后的部分。利用这个功能，可以让AI按照预设的格式续写。

## 自动截断

API服务端会自动截断输入的会话消息文本，从而保证调用AI模型前，有足够空间生成max\_new\_tokens个tokens。

具体来说，截断的目标是要满足如下条件：

TokenLen(System Prompt) + TokenLen(截断后的会话消息文本) + max\_new\_tokens + 50 <= ModelMaxLen

其中，TokenLen表示一个字符串的token数，通常一个汉字/一个英文单词是一个token。ModelMaxLen是模型的最大输入长度，例如2048。

截断会优先丢弃旧的（在数组中位于前面的）消息。截断并不是以一条消息为单位的，也就是说，截断可能会导致某条消息只保留一部分内容，这样可以尽可能多地保留上下文信息。

请注意，当输入的max\_new\_tokens较大，或者输入的聊天历史较长时，将会导致截断的发生。建议合理设置max\_new\_tokens，并在调用请求前丢弃过长的聊天历史（例如，只保留最近10条消息）。

此外，当会话消息总长度超过60000个tokens时，该请求可能会被直接拒绝。

## 响应格式

服务器返回的是流式传输的JSON lines，每一行用\n分隔，都是一个JSON对象。例如：

{"o":"Hello! How can I "}  
{"o":"help you today?\n"}  
{"done":true}

响应对象可能包含以下字段：

* o: 文本片段
* done: 表示生成已完成
* err: 错误信息

## 数据处理

TCP传输过程中数据可能会分开、合并传输，不能假定每次传输的数据是一行或者是一个完整的UTF8字符串。因此，客户端需要将数据缓存到buffer里，同时不断从buffer里查找\n字节，找出并处理完整的行。

## 错误处理

err字段可能会在生成开始时或者生成过程中产生。如果在生成任何文本之前就发生错误，那么就应该按照错误处理。如果生成了一部分文本后发生了错误，那么可以保留已生成的部分，并视为正常结果处理，保存至会话记录并为下一次交互做准备。