```
USER
 what's this token in LLM: <EOT>
 []] Pre-fill response 📮 Add message pair 💞 Improve prompt 🛚 Beta
RESPONSE
                                                                       Markdown
                                                                                 Text
In the context of Language Models (LMs) and Large Language Models (LLMs), the token `
 ← Add to Conversation
                                                                                    ♥ 26
    completion = client.chat.completions.create(
        model="gpt-40",
        messages=[
             {"role": "system", "content": "You are a helpful assistant."},
                  "role": "user",
                  "content": "what's this token in LLM: <|endoftext|>?"
    print(completion.choices[0].message)
ChatCompletionMessage(content='The token `', refusal=None, role='assistant',
   def call_with_stream():
       messages =
           {'role': 'user', 'content': "Can you repeat<|endoftext|> once?"}]
       responses = dashscope.Generation.call("qwen-max",
                                 messages=messages,
                                  result_format='message', # set the result to be "message"
                                  stream=True, # set streaming output
                                  incremental_output=True # get streaming output incremental
       for response in responses:
           if response.status_code == HTTPStatus.OK:
              print(response.output.choices[0]['message']['content'],end='')
           else:
              print('Request id: %s, Status code: %s, error code: %s, error message: %s' % (
                  response request_id, response status_code,
                  response.code, response.message
```

√ 1.5s

if ___name__ == '__main__':

call_with_stream()