**Introduction**

This document is an exploration of how OpenAI logs its requests with language models they host. The reason this was investigated is that we weren’t aware of batch requests at first, and thought there could be an issue in the prompt.

**Content**

How GPT makes requests:

* In the playground, for each time you press the submit button, the OpenAI API makes 1 request.
* But, when calling the API from the code, multiple requests are made at the same time (batch requests), as you can see below.
  + Average batch size of 18.

Few-shot request:

6:36 PM

Local time: Sep 14+ 2022+ 11:36 AM

text-davinci, 5 requests

8+511 prompt + 349 completion = 8+860 tokens

One-shot requests:

11 + 4 + 20 + 22 + 19 + 21 + 17 + 19 + 20 + 19 + 21 + 20 + 20 + 18 + 18 + 20 + 19 + 22 + 20 + 19 + 20 + 18 + (17 - 16) = 408

* 17 - 16 because of keyboard interrupt

11 requests

4 requests

20 requests

22 requests

19 requests

21 requests

17 requests

19 requests

20 requests

19 requests

21 requests

20 requests

20 requests

18 requests

18 requests

20 requests

19 requests

22 requests

20 requests

19 requests

20 requests

18 requests

17 requests