How can I solve 429: 'Too Many Requests' errors?

Updated over 10 months ago

Rate limit errors ('Too Many Requests', ‘Rate limit reached’) are caused by hitting your organization's [rate limit](https://help.openai.com/en/articles/5955598) which is the maximum number of requests and tokens that can be submitted per minute. If the limit is reached, the organization cannot successfully submit requests until the rate limit is reset. The error message looks like this:

Rate limit reached for gpt-3.5-turbo in organization org-exampleorgid123 on tokens per min.   
Limit: 10000.000000 / min. Current: 10020.000000 / min.

We recommend handling these errors using exponential backoff. Exponential backoff means performing a short sleep when a rate limit error is hit, then retrying the unsuccessful request. If the request is still unsuccessful, the sleep length is increased and the process is repeated. This continues until the request is successful or until a maximum number of retries is reached.

As unsuccessful requests contribute to your per-minute limit, continuously resending a request won’t work. Rate limits can be applied over shorter periods - for example, 1 request per second for a 60 RPM limit - meaning short high-volume request bursts can also lead to rate limit errors. Exponential backoff works well by spacing apart requests to minimize the frequency of these errors.

In Python, an exponential backoff solution could look like this:

from openai.error import RateLimitError  
import backoff  
  
@backoff.on\_exception(backoff.expo, RateLimitError)  
def completions\_with\_backoff(\*\*kwargs):  
response = openai.Completion.create(\*\*kwargs)  
return response

(Please note: The backoff library is a third-party tool. We encourage all our customers to do their due diligence when it comes to validating any external code for their projects.)

If implementing exponential backoff still results in this error, you might need to [increase your usage tier](https://platform.openai.com/docs/guides/rate-limits/usage-tiers). You can view your current rate limits and how to increase your trust tier to increase your rate limits in the [limits section](https://platform.openai.com/account/limits) of your account settings.