Capacity and Performance

How many users can use our system at a time?

If this project were to be produced at an enterprise scale, we would design for scalability, however because this is a local project, scalability is not a business concern. The numbers should be dependent on the capacity of the Kent State servers.

The capacity of our system will be determined by the number of users supported by the web server, the number of requests sent to the API, and the average response time from the API.

Depending on the API model used, one token cycle (input tokens and output tokens for a prompt and response) would be approximately 70 tokens, for which the price would vary. There is not a token limit for the API. The average response time for GPT 4o mini API is 196 milliseconds per generated token.

As the number of requests increase, and the API scales its requests, latency is generally predicted to increase, and performance is generally predicted to decrease. Input and output length will affect latency and performance.

server:
-kent state server: can support ___ users at a time

limitations of API: GPT 4o mini
-number of requests: unlimited
-average response time per request: 196 milliseconds per generated token

GPT-3.5 and GPT-4 API response time measurements - FYI - API - OpenAI Developer Forum