# Performance Test Report - Mar 8, 2025 (#2)

Open in Postman

Postman collection: Azure OpenAI Service API

Report exported on: Mar 8, 2025, 22:19:52 (GMT+5:30)

## Test setup

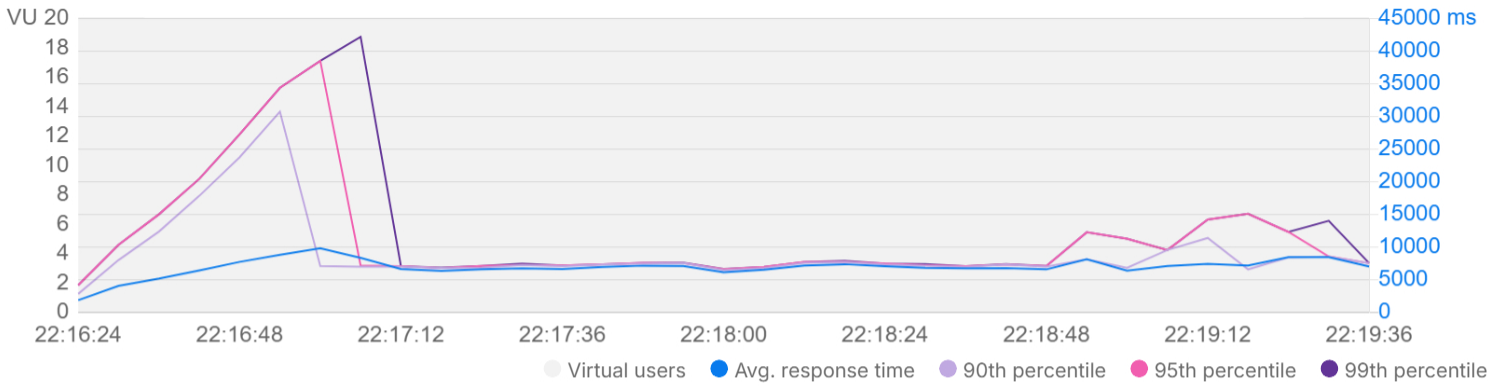
| Virtual users | Start time | Load profile |
|---|---|---|
| 20 VU | Mar 8, 22:16:21 (GMT+5:30) | Fixed |
| **Duration** | **End time** | **Environment** |
| 10 minutes (Terminated at 3 minutes 21 seconds) | Mar 8, 22:19:43 (GMT+5:30) | - |

## 1. Summary

| Total requests sent | Throughput | Average response time | Error rate |
|---|---|---|---|
| 479 | 2.38 requests/second | 6,966 ms | 0.63 % |

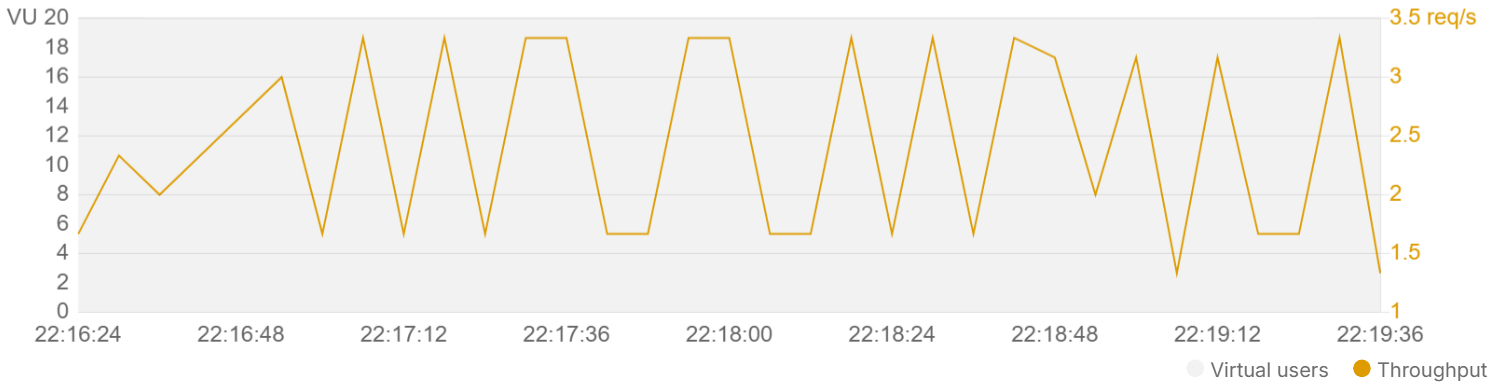
### 1.1 Response time

Response time trends during the test duration.




### 1.2 Throughput

Rate of requests sent per second during the test duration.



### 1.3 Requests with slowest response times

Top 5 slowest requests based on their average response times.

| Request | Resp. time (Avg ms) | 90th (ms) | 95th (ms) | 99th (ms) | Min (ms) | Max (ms) |
|---|---|---|---|---|---|---|
| POST local chat compose Copy 2<br>http://localhost:8000/openai/deployments/:deployment-id/chat/completions?api-version=2024-02-15-preview | 6,966 | 7,781 | 8,451 | 27,275 | 673 | 42,190 |

### 1.4 Requests with most errors

Top 5 requests with the most errors, along with the most frequently occurring errors for each request.

| Request | Total error count | Error 1 | Error 2 | Other errors |
|---|---|---|---|---|
| POST local chat compose Copy 2<br>http://localhost:8000/openai/deployments/:deployment-id/chat/completions?api-version=2024-02-15-preview | 3 | ECONNRESET (3) | - | 0 |

## 2. Metrics for each request

The requests are shown in the order they were sent by virtual users.

| Request | Total requests | Requests/s | Min (ms) | Avg (ms) | 90th (ms) | Max (ms) | Error % |
|---|---|---|---|---|---|---|---|
| POST local chat compose Copy 2<br>http://localhost:8000/openai/deployments/:deployment-id/chat/completions?api-version=2024-02-15-preview | 479 | 2.38 | 673 | 6,966 | 7,781 | 42,190 | 0.63 |

## 3. Errors

### 3.1 Error distribution over time

Top 5 error classes observed during the test duration.


### 3.2 Error distribution for requests

Errored requests grouped by error class, along with the error count for each class.

| Error class | Total counts |
|---|---|
| ECONNRESET | 3 |
| POST local chat compose Copy 2 | 3 |

### Testing API performance on Postman

Postman enables you to simulate user traffic and observe how your API behaves under load. It also helps you identify any issues or bottlenecks that affect performance.

Learn more about testing API performance.