LLM Performance Benchmark Report

# Introduction

This report presents the results of performance benchmarking for Large Language Models (LLMs). The benchmarks measure tokens per second, latency, and throughput under various concurrency levels.

# Test Configuration

## Models

|  |  |
| --- | --- |
| Model Name | Base URL |
| gpt-3.5-turbo | https://api.openai.com/v1 |

## Test Cases

|  |  |
| --- | --- |
| Test Name | Input Prompt |
| short\_prompt | Explain the concept of machine learning in one paragraph. |
| medium\_prompt | Write a short essay about the impact of artificial intelligence on society. |

## Test Parameters

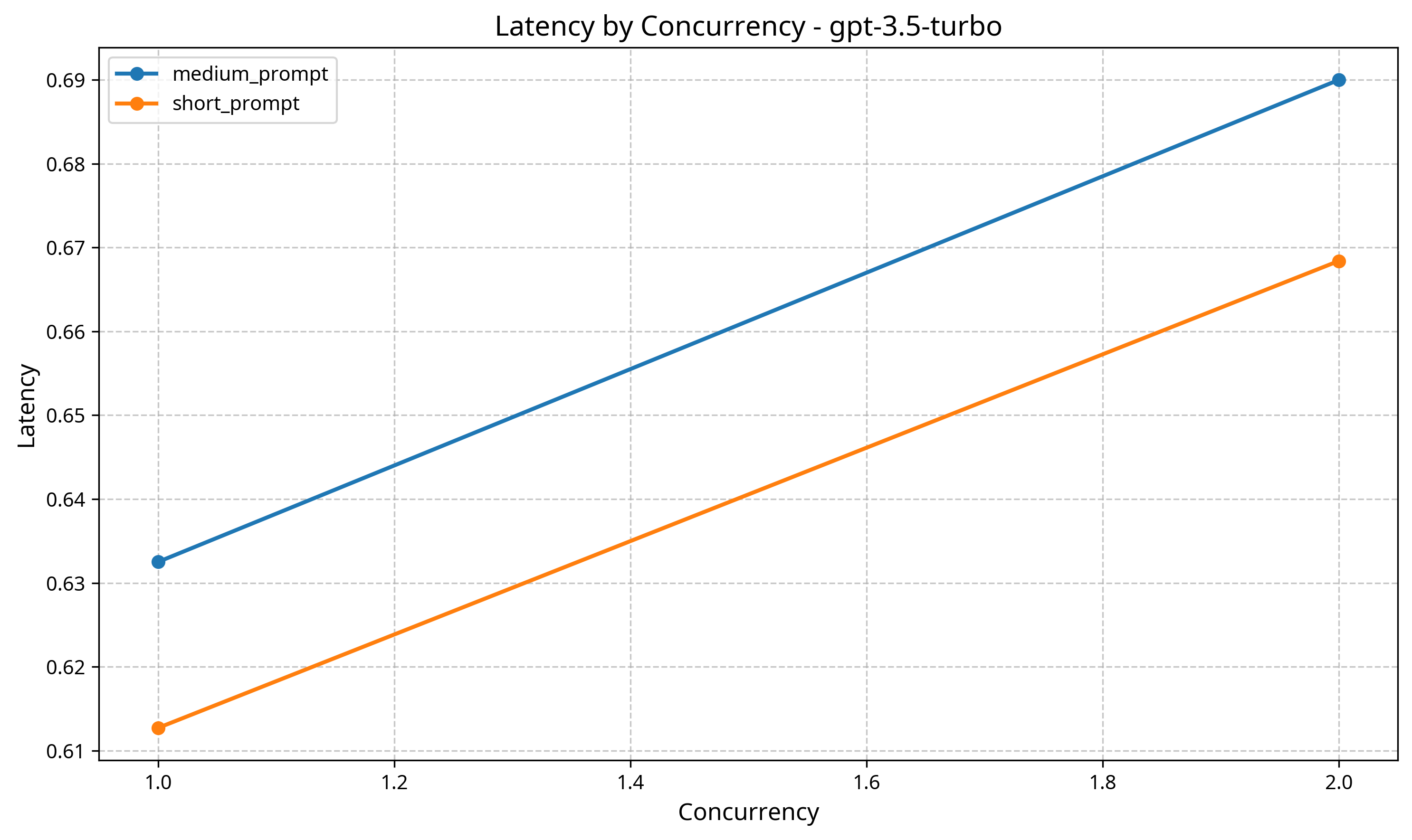
Concurrency Levels: 1, 2

Total Requests per Test: 5

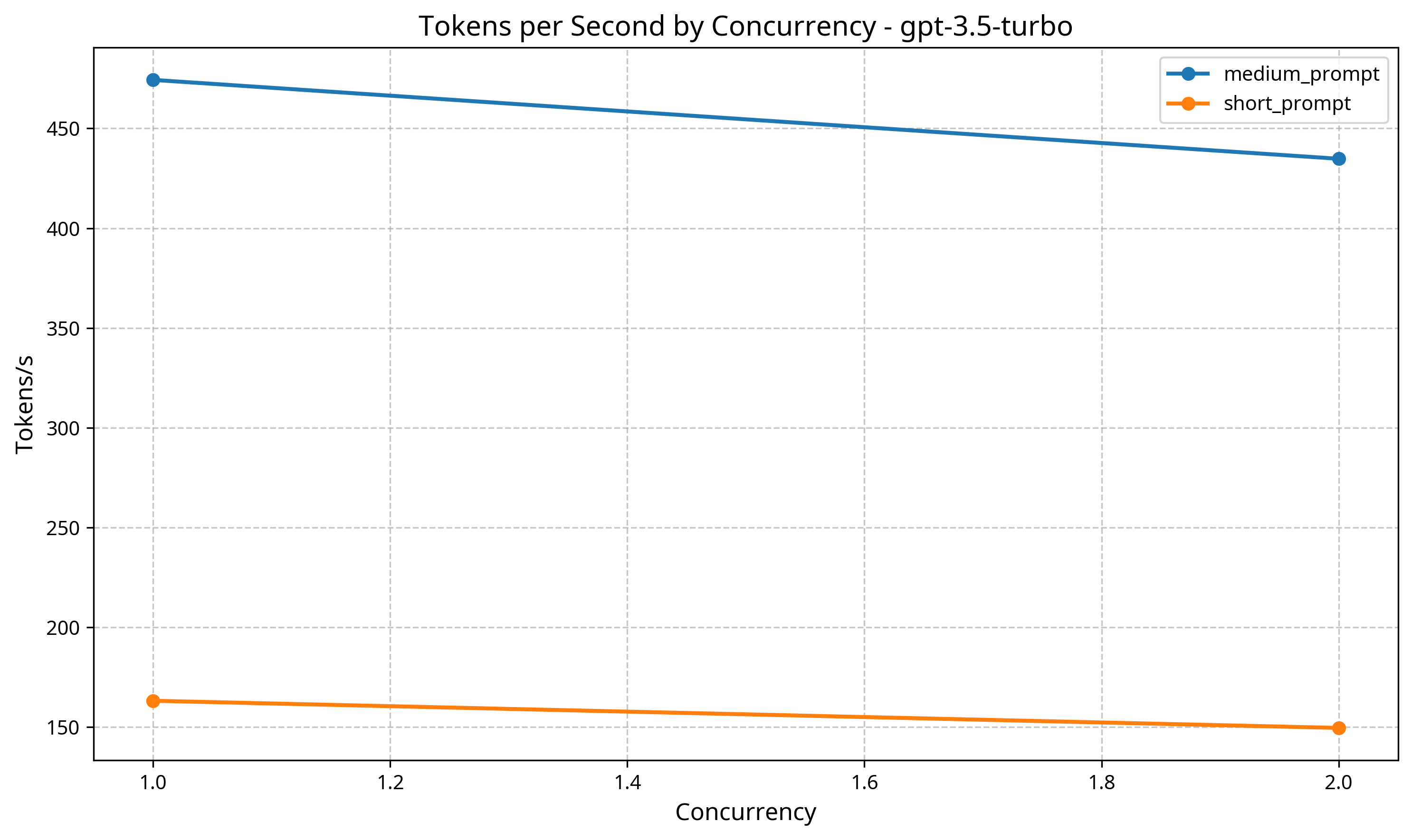
# Performance Results

## Model: gpt-3.5-turbo

### Latency by Concurrency



### Tokens per Second by Concurrency



### Test: short\_prompt

|  |  |  |  |
| --- | --- | --- | --- |
| Concurrency | Avg. Latency (s) | Tokens/s | Success Rate |
| 1 | 0.613 | 163.21 | 100.0% |
| 2 | 0.668 | 149.61 | 100.0% |

### Test: medium\_prompt

|  |  |  |  |
| --- | --- | --- | --- |
| Concurrency | Avg. Latency (s) | Tokens/s | Success Rate |
| 1 | 0.632 | 474.31 | 100.0% |
| 2 | 0.690 | 434.78 | 100.0% |

# Automated Analysis

## General Observations

* Average latency across all tests: 0.651 seconds.
* Average tokens per second across all tests: 305.48.
* Average success rate across all tests: 100.0%.

## Concurrency Impact

* Latency remains relatively stable (change of 9.1%) across concurrency levels.
* Tokens per second remains relatively stable (change of -8.3%) across concurrency levels.
* The optimal concurrency level for maximum tokens per second appears to be 1.

## Recommendations

* Consider running longer tests with more requests to get more statistically significant results.
* For optimal throughput, consider using a concurrency level of 1.
* When sizing infrastructure, ensure that you account for peak load scenarios by adding a buffer of at least 30% to the average throughput requirements.
* Monitor both latency and tokens per second metrics in production, as they may vary based on input complexity and output length.