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

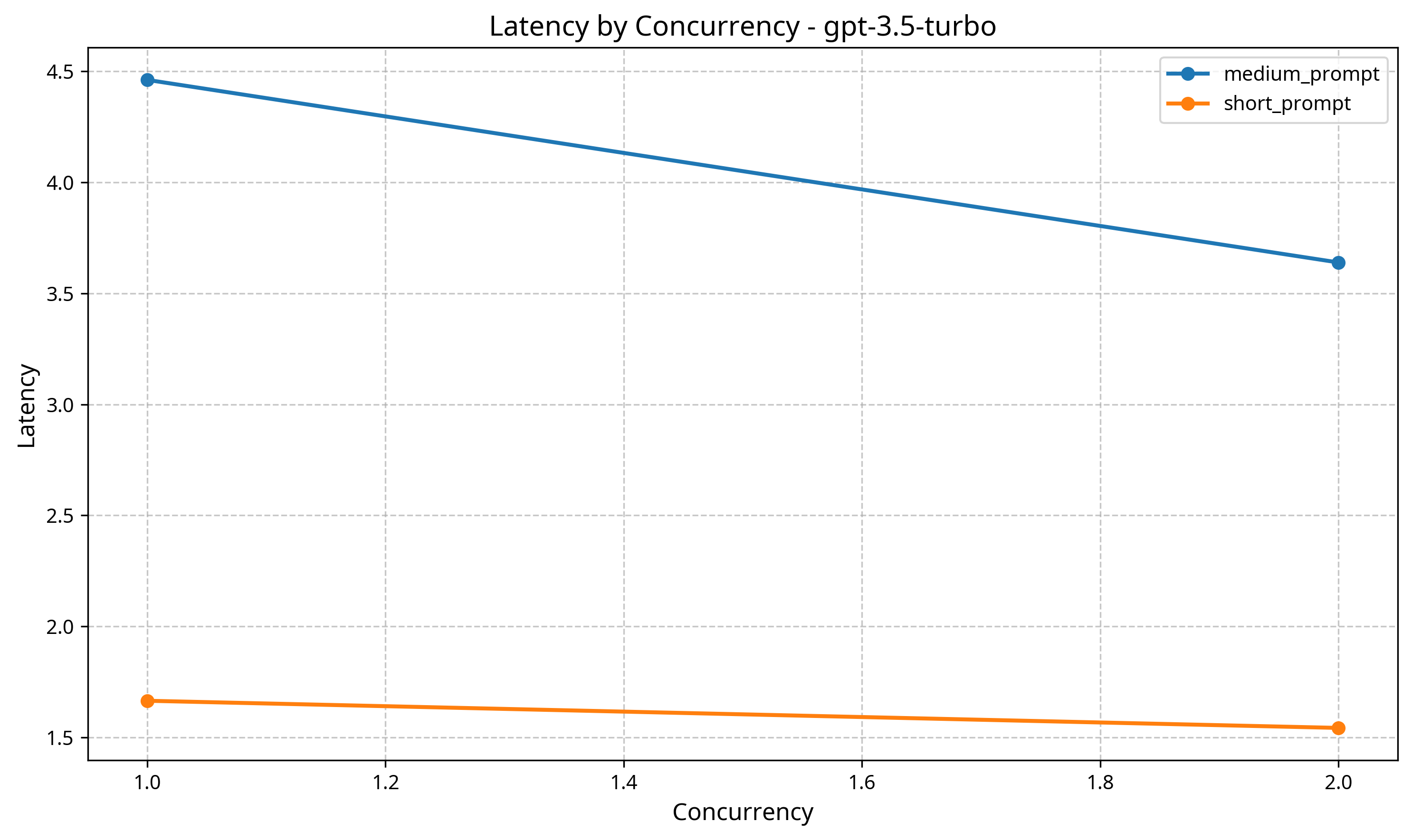
Concurrency Levels: 1, 2

Total Requests per Test: 5

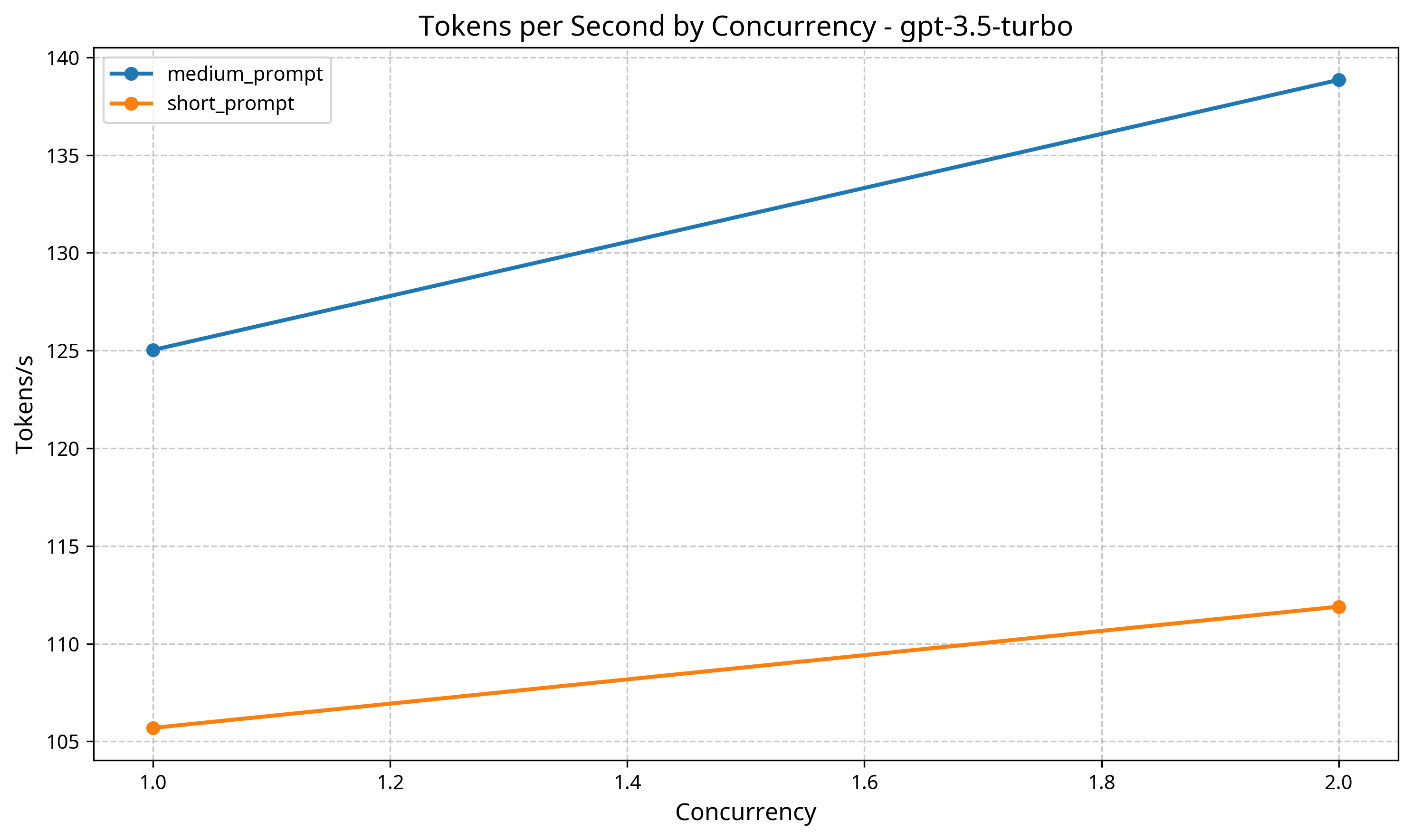
# Performance Results

## 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 | 1.665 | 105.69 | 100.0% |
| 2 | 1.543 | 111.90 | 100.0% |

### Test: medium\_prompt

|  |  |  |  |
| --- | --- | --- | --- |
| Concurrency | Avg. Latency (s) | Tokens/s | Success Rate |
| 1 | 4.462 | 125.03 | 100.0% |
| 2 | 3.639 | 138.85 | 100.0% |

# Automated Analysis

## General Observations

* Average latency across all tests: 2.827 seconds.
* Average tokens per second across all tests: 120.36.
* Average success rate across all tests: 100.0%.

## Concurrency Impact

* Latency decreases by 15.4% when concurrency increases from 1 to 2, which is unusual and may indicate caching effects or warm-up benefits.
* Tokens per second remains relatively stable (change of 8.7%) across concurrency levels.
* The optimal concurrency level for maximum tokens per second appears to be 2.

## Recommendations

* Consider running longer tests with more requests to get more statistically significant results.
* For optimal throughput, consider using a concurrency level of 2.
* Since the highest tested concurrency level (2) provided the best throughput, consider testing even higher concurrency levels to find the true optimal point.
* 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.