# Production Metrics

vLLM exposes a number of metrics that can be used to monitor the health of the system. These metrics are exposed via the `/metrics` endpoint on the vLLM OpenAI compatible API server.

You can start the server using Python, or using [Docker](#deployment-docker):

```console
vllm serve unsloth/Llama-3.2-1B-Instruct
```

Then query the endpoint to get the latest metrics from the server:

```console
$ curl http://0.0.0.0:8000/metrics

# HELP vllm:iteration_tokens_total Histogram of number of tokens per engine_step.
# TYPE vllm:iteration_tokens_total histogram
vllm:iteration_tokens_total_sum{model_name="unsloth/Llama-3.2-1B-Instruct"} 0.0
vllm:iteration_tokens_total_bucket{le="1.0",model_name="unsloth/Llama-3.2-1B-Instruct"} 3.0
vllm:iteration_tokens_total_bucket{le="8.0",model_name="unsloth/Llama-3.2-1B-Instruct"} 3.0
vllm:iteration_tokens_total_bucket{le="16.0",model_name="unsloth/Llama-3.2-1B-Instruct"} 3.0
vllm:iteration_tokens_total_bucket{le="32.0",model_name="unsloth/Llama-3.2-1B-Instruct"} 3.0
vllm:iteration_tokens_total_bucket{le="64.0",model_name="unsloth/Llama-3.2-1B-Instruct"} 3.0
vllm:iteration_tokens_total_bucket{le="128.0",model_name="unsloth/Llama-3.2-1B-Instruct"} 3.0
vllm:iteration_tokens_total_bucket{le="256.0",model_name="unsloth/Llama-3.2-1B-Instruct"} 3.0
```

```
vllm:iteration_tokens_total_bucket{le="512.0",model_name="unsloth/Llama-3.2-1B-Instruct"} 3.0
...
```

The following metrics are exposed:

:::{literalinclude} ../../../vllm/engine/metrics.py

:end-before: end-metrics-definitions

:language: python

:start-after: begin-metrics-definitions

:::