Environment Variables

vLLM uses the following environment variables to configure the system:

:::{warning}
Please note that `VLLM_PORT` and `VLLM_HOST_IP` set the port and ip for vLLM's **internal usage**. It is not the port and ip for the API server. If you use `--host $VLLM_HOST_IP` and `--port $VLLM_PORT` to start the API server, it will not work.

All environment variables used by vLLM are prefixed with `VLLM_`. **Special care should be taken for Kubernetes users**: please do not name the service as `vllm`, otherwise environment variables set by Kubernetes might conflict with vLLM's environment variables, because [Kubernetes sets environment variables for each service with the capitalized service name as the prefix](https://kubernetes.io/docs/concepts/services-networking/service/#environment-variables).
:::

:::{literalinclude} ../../../vllm/envs.py
:end-before: end-env-vars-definition
:language: python
:start-after: begin-env-vars-definition
:::