```
# serve with open ai api
envs:
  MODEL_NAME: meta-llama/Llama-2-7b-chat-hf
  HF_TOKEN: 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

resources:
  accelerators: {A1000}
  ports:
    - 8000

setup: |
  conda activate vllm
  if [ $? -ne 0 ]; then
    conda create -n vllm python=3.10 -y
    conda activate vllm
  fi

  pip install transformers==4.38.0
  pip install vllm==0.3.2
  python -c "import huggingface_hub; huggingface_hub.login('${HF_TOKEN}')"


run: |
  conda activate vllm
  echo 'Starting vllm openai api server...'
  python -m vllm.entrypoints.openai.api_server \
```

```
--model $MODEL_NAME --tokenizer hf-internal-testing/llama-tokenizer \

--host 0.0.0.0
```