SIDE PROJECT

AI server for high-performance, local

Mistral

Running the Docker Container

First, ensure you have Docker installed and your GPU drivers are up to date. Use the following command to run the Mistral AI LLM Inference image:

```
zafar@auth-srv:~$ docker run --gpus all \
    -e HF_TOKEN=$HF_TOKEN -p 8000:8000 \
    ghcr.io/mistralai/mistral-src/vllm:latest \
    --host 0.0.0.0 \
    --model mistralai/Mistral-7B-Instruct-v0.2
Unable to find image 'ghcr.io/mistralai/mistral-src/vllm:latest' locally
latest: Pulling from mistralai/mistral-src/vllm
43f89b94cd7d: Pull complete
45f7ea5367fe: Pull complete
3d97a47c3c73: Pull complete
12cd4d19752f: Pull complete
da5a484f9d74: Pull complete
5e5846364eee: Downloading [=======>                                           ]  163.4MB/1.291GB
fd355de1d1f2: Download complete
3480bb79c638: Download complete
e7016935dd60: Download complete
99541166a133: Downloading [>                                                  ]  35.63MB/2.509GB
8999112df5b0: Download complete
e969c5eb17ee: Download complete
174617b6ae76: Download complete
7fcb0eeb3246: Waiting
8546325b89a2: Waiting
fd3e44b6510f: Waiting
1ad8795b31a4: Waiting
962181193532: Waiting
ccb00ad5abb9: Waiting
fa4989232485: Waiting
```