DrDroid

VLLM Cheatsheet

Connection

curl -X POST http://{server_address}:{port}/v1/completions -H "Content-Type: application/json" -d '{"prompt": "Hello", "max_tokens": 100}'
Connect to vLLM server

python -m vllm.entrypoints.api_server --model {model_name} --port {port}
Set up vLLM server

Basic Operations

curl -X POST http://{server_address}:{port}/v1/completions -H "Content-Type: application/json" -d '{"prompt": "Your prompt here", "max_tokens": 100, "temperature": 0.7}'
Text completion

curl -X POST http://{server_address}:{port}/v1/chat/completions -H "Content-Type: application/json" -d '{"messages": [{"role": "user", "content": "Hello"}], "max_tokens": 100}'
Chat completion

curl -X POST http://{server_address}:{port}/v1/completions -H "Content-Type: application/json" -d '{"prompt": "Your prompt", "max_tokens": 100, "stream": true}'
Streaming completion

Engine Management

python -m vllm.entrypoints.api_server --model {model_name}
Load model

python -m vllm.entrypoints.api_server --model {model_name} --gpu-ids 0,1,2,3
Specify GPU devices

python -m vllm.entrypoints.api_server --model {model_name} --tensor-parallel-size 4
Set tensor parallelism

Inference Settings

Add "temperature": 0.7 in request JSON
Set temperature

Add "top_p": 0.95 in request JSON
Set top-p sampling

Add "max_tokens": 512 in request JSON
Set maximum output length

Add "presence_penalty": 1.0 in request JSON
Set repetition penalty

Batch Processing

python -m vllm.entrypoints.api_server --model {model_name} --max-model-len {length} --max-num-batched-tokens {num_tokens}
Set maximum batch size

curl -X POST http://{server_address}:{port}/v1/completions -H "Content-Type: application/json" -d '{"prompt": ["Prompt 1", "Prompt 2"], "max_tokens": 100}'
Multi-prompt batching

Monitoring

curl http://{server_address}:{port}/health
Check server status

curl http://{server_address}:{port}/metrics
Get server metrics