

Inference Broker Web

Chat

Embedding

OpenAI Chat Completions

OpenAI Responses

OpenAI Embedding

Whisper Transcribe

Describe

Image Generation

Image Img2img

Inference Broker Information

Connection

Status

Connected

Base URL

http://host.docker.internal:21000

Health

ok

Local

Runtime Agent

http://127.0.0.1:21000

Backends

Backend	Kind	Model	Runtime	Status	Control
text-main	text_generation	qwen	llamacpp	running	
embedding-main	embedding	qwen	llamacpp	running	
text-diffusion	text_generation	qwen	diffusers	stopped	

GPU

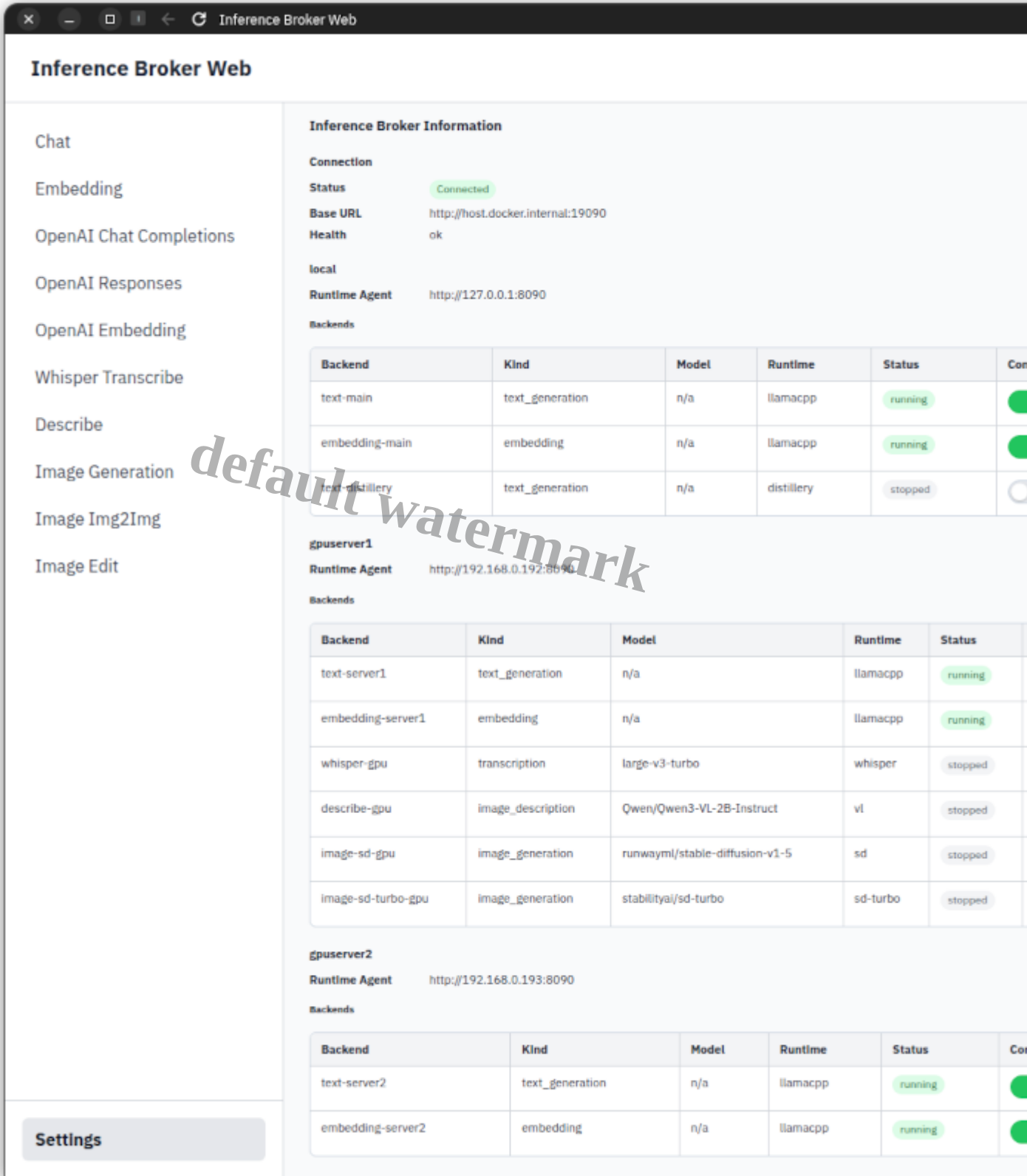
Description

default watermark

ã?•ã??ã?¬ã?°ã?©ã?|ã?³ã??ç³»ã•®æ?¬è«?ã?|ç?•ã??ç?»ã?•ç³»ã?|ç?•ã?«ã½¿ã•£ã•|ã•
?ã??GPUã?µã?¼ã?•ã?¼i¼?dGPUã??2æ??æ•è¼?i¼?ã•®ã»?ã«ã?•ã³¼è©±ç³»æ?¬è«?ã?|ç?ã•§ã½¿ã
•?GPUã?µã?¼ã?•ã?¼ã??ç?¬æ?•ã?ã•¾ã?ã?ã??

VRAMã•®ã±§ã•ã•GPUã??ã ¥æ??ã?ã??ã•¾ã•§ã•®æ?«ã®?æ?ªç½®ã•ã•®ã•§ã?•2ã°ç?®ã•
®ã?µã?¼ã?•ã?¼ã•®GPUã•ã•ã??ã?ã?ã?ã?ã?ã?ã•RTX 3060 12GBã•é•?ã??ã•§ã?ã??

ä,?ã•®ã?¹ã?¬ã?ªã?¼ã?³ã?•ã?§ã??ã??ã¬é??ç?°æ©?ã•®Inference Broker UIã•§ã?ã?ã?•gpuserver1ã•
?ã?ã??ã•¾ã•§ç¼¼ã?•ã?ã?ã?ã?ã?ã?¹ã°ç?®ã•®GPUã?µã?¼ã?•ã?¼ã•§ã?•gpuserver2ã•
?ã»?æ?¥è¿½ã?ã?ã?GPUã?µã?¼ã?•ã?¼ã•§ã?ã??

[illegible]

ã•?ã•ã•?ã•£ã•|ã•?ã•,æ??ç¼¼ã•?ã•?ã•?ã•?ã•µã¼ã•?ã•¼ã•ªã•®ã•§ã•?ã•?ã•¼ã•?æ¶?è²»é?»ã•?ã•
®ã•ªã•ã•ªGPUã•¯é¿ã•?ã•?ã•?ã•?ã•?ã•?ã•?
CUDAã•?³ã•?¢æ°ã•?è¨?ç®?é??ã•!ã•?ã•?ã•?ã•?ã•VRAMã•®¹é?ã•?ã•?ã•?ã•?ã•?ã•?ã•?
RTX 3060ã•«24GBã•ã•?ã•?ã•?ã•®VRAMã•?è¼ã•?£ã•|ã•?ã•?ã•?ã•?ã•?ã•?ã•®ã•«ã•£ã•|ã•?ã•ªã•?æ?ã•?ã•¼ã•
?ã•?
3060ã•§ã•?llama.cpp + gpt-oss:20b Q4ã•§ã•?ã•?¼ã•?ã•³ç??æ?é??ã•!ã•¯100tpsè¶ã•?ã•?ã•?ã•?
ã•³ã•?ã•?ã•?ã•?|ç•?ã•¯300tpsè¶ã•?ã•§ã•?ã•?ã•?ã•®ã•ã•?ã•?ã•?ã•æ?§è?½ç??ã•«ã•?ã•?ã•?ã•ã•?ã•?ã•
§ã•ã•?ã•?

Category

1. æ?¥ã•?ã•®æ´»ã•?

Tags

1. AI
2. gpt-oss
3. Inference Broker
4. llama.cpp
5. LLM
6. ä_?é??ã•?ã•?
7. å±±±æ¢¨ç??

Date Created

2026å¹10æ??31æ?¥

Author

kazuo-tsubaki