

Inference Broker â?? GPUâ??â½¿ã•£ã•?Visionç³»ã?¢ã??ã?«ã•,ã•®ã³¼ã¿

Description

ã•?ã•?ã•¼ã•§ã•?Inference Brokerã•?ã³¼ã¿?ã•?ã•?ã•?æ?°è«?è³?æ°ã•ã•?

- llama.cpp Vulcanã•?ã??ã?ã?ã?ã??
- AMD Stable Diffusion
- Lemonadeã??ã½¿ã•£ã•?Whisper
- NPUã??ã½¿ã•£ã•?Stable Diffusion
- Lemonadeã??ã½¿ã•£ã•?Stable Diffusion
- Lemonadeã??ã½¿ã•£ã•?Visionã?¢ã??ã?«
- llama.cpp Cudaã•?ã??ã?ã?ã?ã??

ã•§ã•?ã•?ã•?ã•?æ?°ã•?ã•«ã•?

- Cudaç??Visionã?¢ã??ã?«i¼?Vision Brokerã•®æ?ã½?ã•ã•£ã•?i¼?
- Cudaç??Stable Diffusioni¼?Vision Brokerã•®æ?ã½?ã•ã•£ã•?i¼?
- Cudaç??Whisperi¼?æ?°i¼?

ã•«ã??ã³¼ã¿?ã•?ã•¼ã•?ã??

ã•?ã•®ã³¼ã¿?ã?çµ?ã•?ã??ã•ã•?Inference Brokerã•?NVIDIA GPU/Ryzen AI CPU/NPUã??ã½¿ã•£ã•?æ?°è«?è³?æ°ã•?ã•,æ??ã•«ã¼?ã•ã•?ã•?ã•?ã•?ã•?ã•§ã•?ã•?LLM Brokerã•?Vision Brokerã•®æ©?è?½ã??ã®?ãã•«ã•,ã•?ã•§ã•?ã•¼ã•?ã??

ä»?æ?¥ã•®é??ç?°ã•§ã•?

- GPU Inferenceã•?ã•?ã•?ã•?Vision/StableDiffusion/Whisperã??GPUä,ã•§ã?©ç?°ã•?ã??ã?µã?¼ã?ã•?¼ã•®é??ç?°
- Inference Runtime Agentã•®æ?jã¼µ

ã??ã??ã??ã•¼ã•?ã•?ã??

ã•?ã??ã•?ZIKUUã•®æ?°è«?ã±ã•®ã®?æ?ã½¿¢ã•§ã•?ã??

Inference Broker

â?? Inference Runtime Agent (ã??ã?¼ã?¢ã?³i¼?

â?? GPU Inference (VL/Stable Diffusion/Whisper)

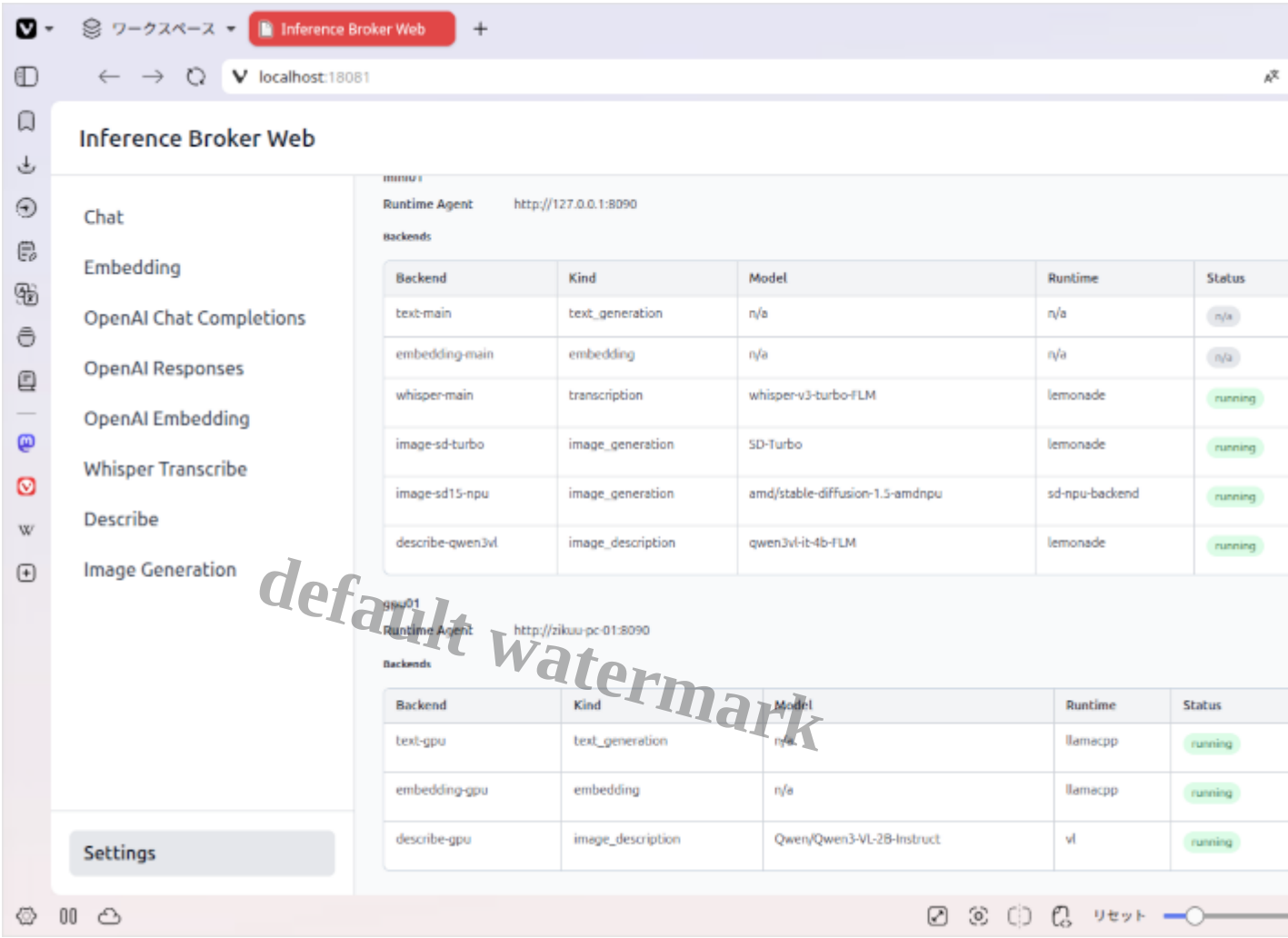
â?? SD NPU Backend (ã??ã?¼ã?¢ã?³/NPUç??Stable Diffusion)

â?? Lemonade

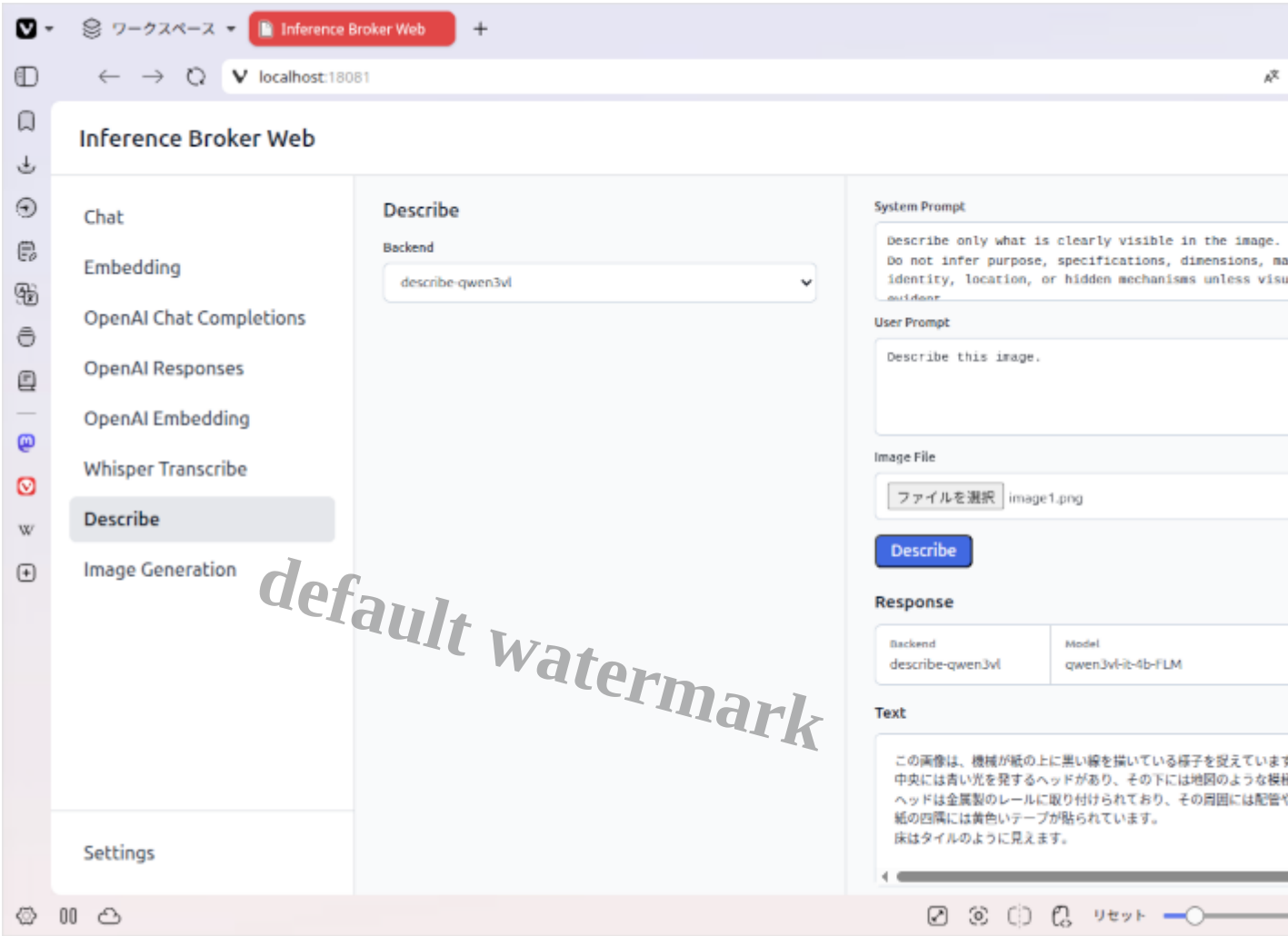
â?? llama.cpp

â?? Ollama

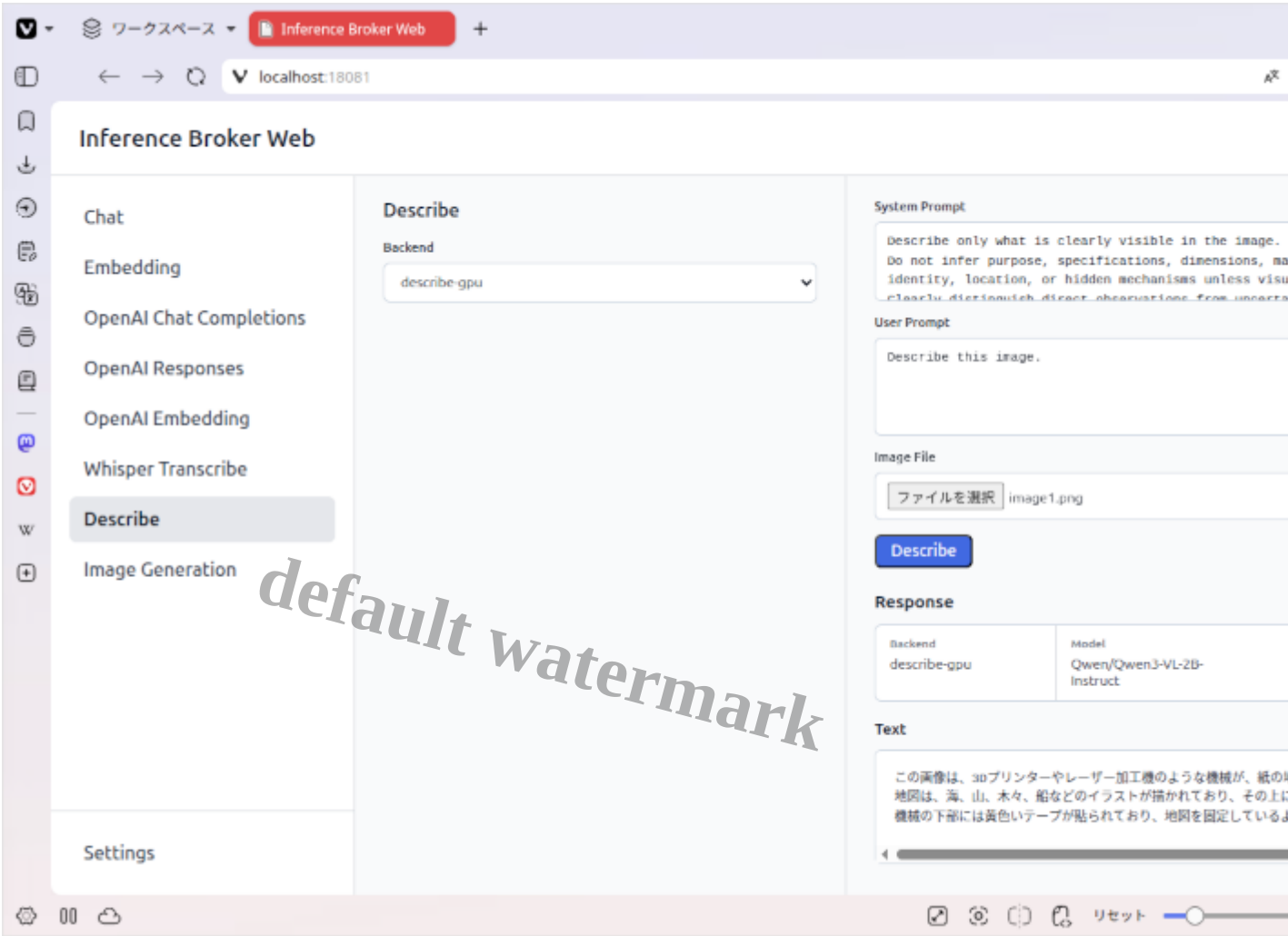
ã•?ã•?ã•?æ§?æ?°ã•§ã•?ã??



ã•ã?ã?Visionã?¢ã??ã?«ã•@ã??ã?¹ã??çµ•æ??ã??
NPUç??



GPU??



Category

1. æ?¥ã? ã•®æ´»å??

Tags

1. AI
2. Inference Broker
3. Inference Runtime Agent
4. LLM
5. ä,?é???å??å,?
6. å±±æ¢ç??

Date Created

2026å¹´9æ??1æ?¥

Author

kazuo-tsubaki