



Inference Broker â?? æ?`è«?è³?æ°•â?•ã??ã?©ã??ã??ã??ã?©ã?¼ã? â??

Description

NPUæ·è¼¼?ã·®PCã·?â? â?·ã·£ã·?ã·?ã·?ã·§ã·?ã·?ã·??ã·¾ã·§ã·®CPU/GPUã·??ã½¿ã·?æ?`è«?â?!ç·
?ã»¥â¤?ã·®é· æ??è?¢ã·?â¢?ã·?ã·¾ã·?ã·?ã·??

ä•ä•æ©?ä¼?ä•«ä?ä•ä??ä•¼ä•sé??ç?°ä•?ä•|ä•ä•?LLM Brokerä•“Vision Brokerä•??çµ±å•?ä•?ä•|ä?
 æ?“è«?è³?æ°•ä•Rç®|ç?ä•?ä•Sä•ä•??æ?“è«?ä•?ä•?©ä•?ä•?ä•?ä•?©ä•?¼ä? ä•«è²ä•|ä•|ä•?ä•?ä•?ä•
 “æ?ä•?ä•¼ä•?ä•??

ä»?æ?¥ã•ã•ã•®é??ç?°ã?ä,?æ®µè½ã?ã¾ã?ã?ã??

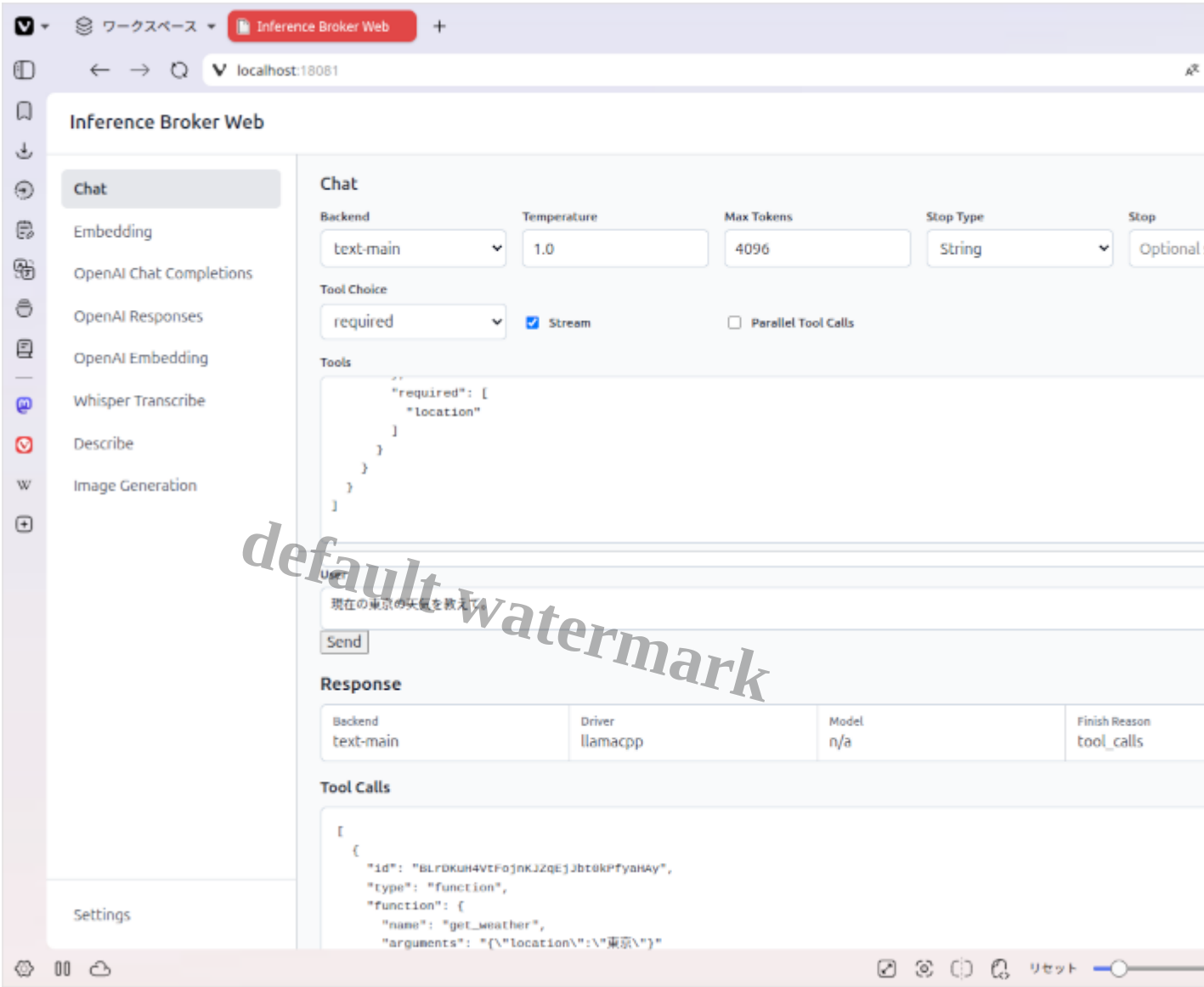
Inference Broker

ä»?æ?¥é??ç?°ã•?ã•?ã•®ã•-ã•?Inference Brokerã•®æ?è"¼ã•"ç®¼ç?ã??ã•?ã??ã•?ã?ã•
®Webã??ã?¼ã?«ã•\$ã•?ã??

ã•?ã??ã•sã»?ã¾4?é??ç?°ã??é?²ã?•ã??é•?ç?`ã•sã?•Inference Brokerã•®æ©?è?½ã•®ã??ã?¹ã??ã??ã•?
?ã•¾4ã•?ã??

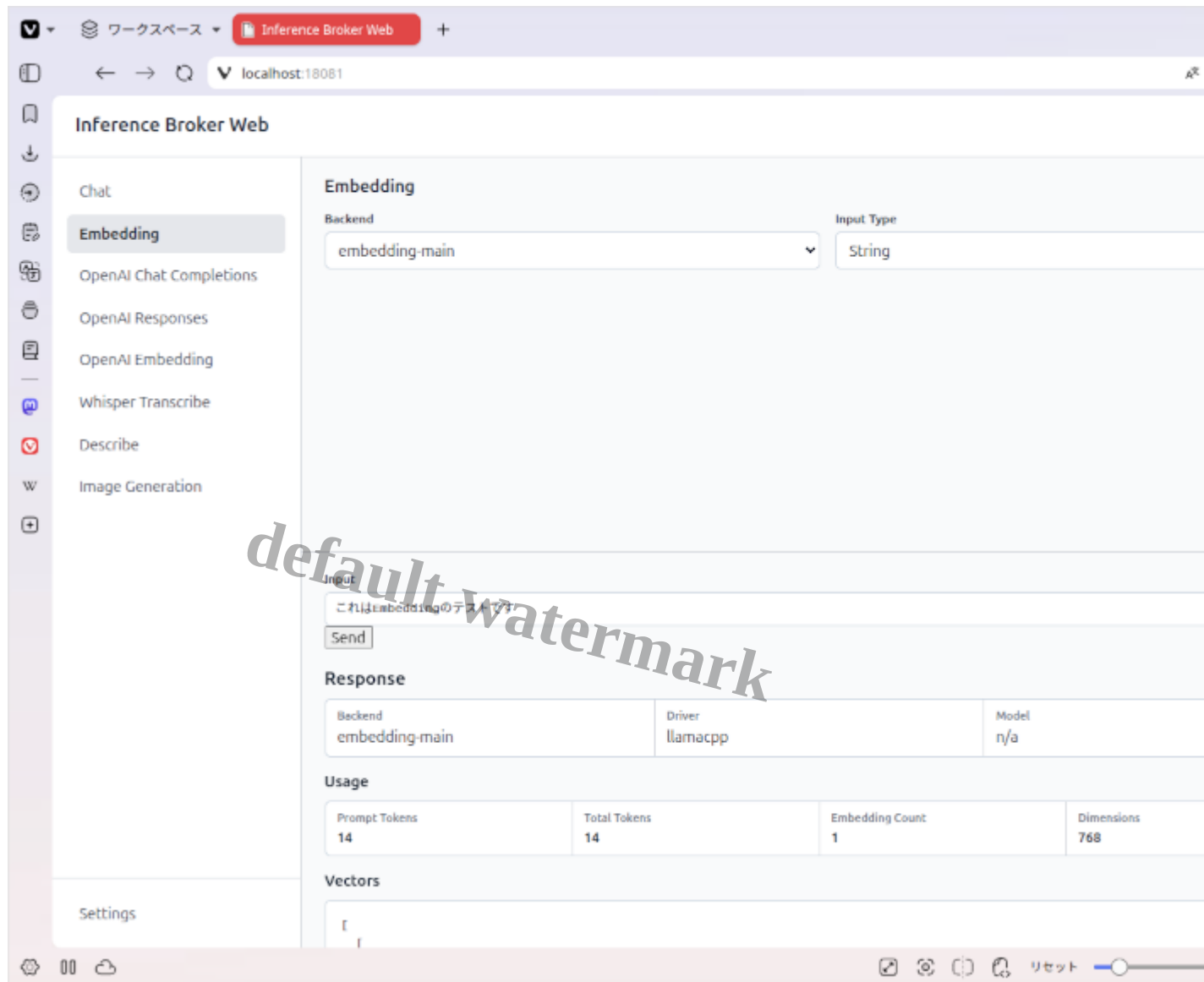
1. Chat

[illegible]



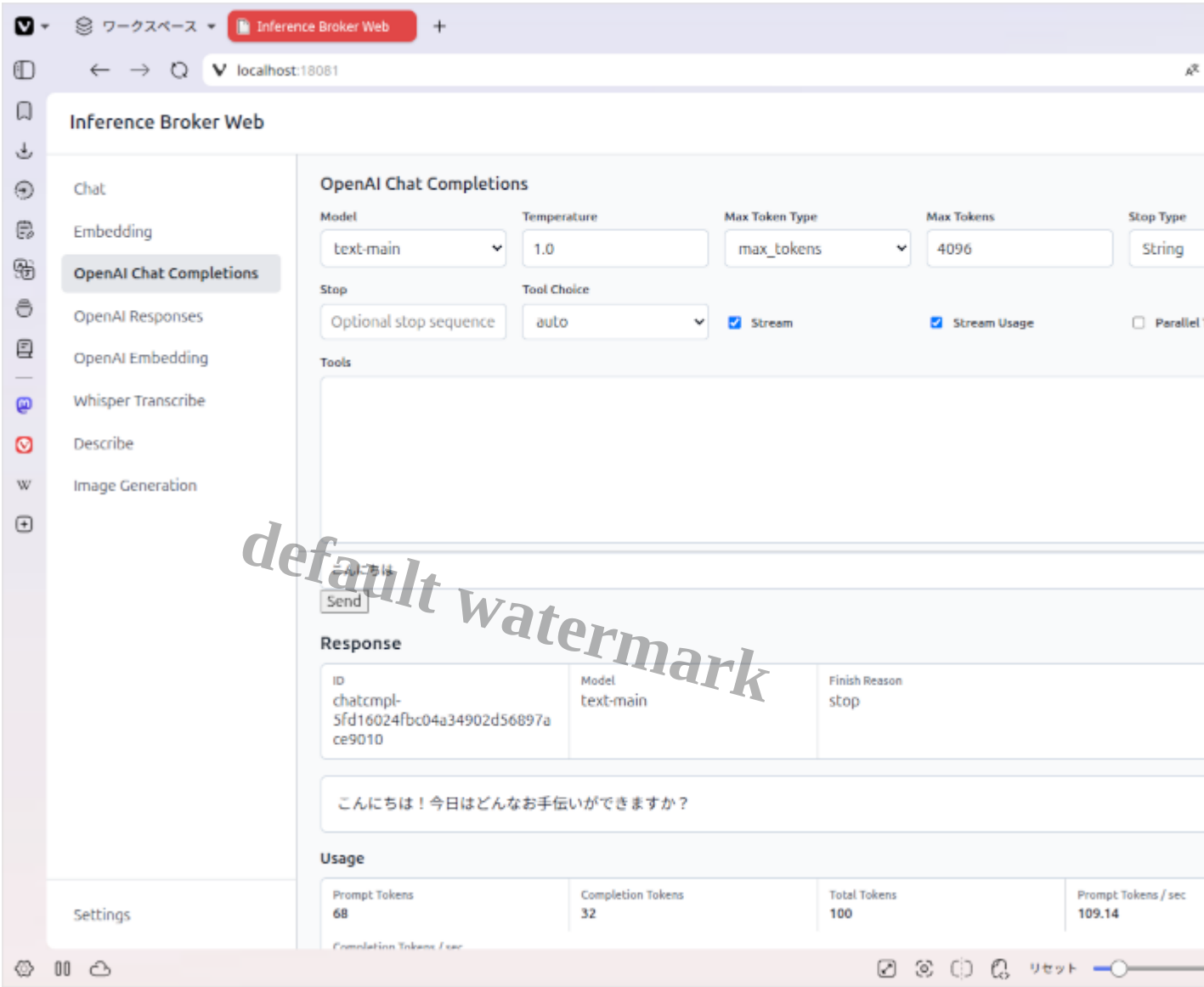
2. Embedding

å??ã?•è¼¼¼ã•¿ã?¢ã??ã?«ã??ä½¿ã•£ã?ã?•å??ã?•è¼¼¼ã•¿ã??ç??æ?•ã?ã??APIã•§ã?ã??



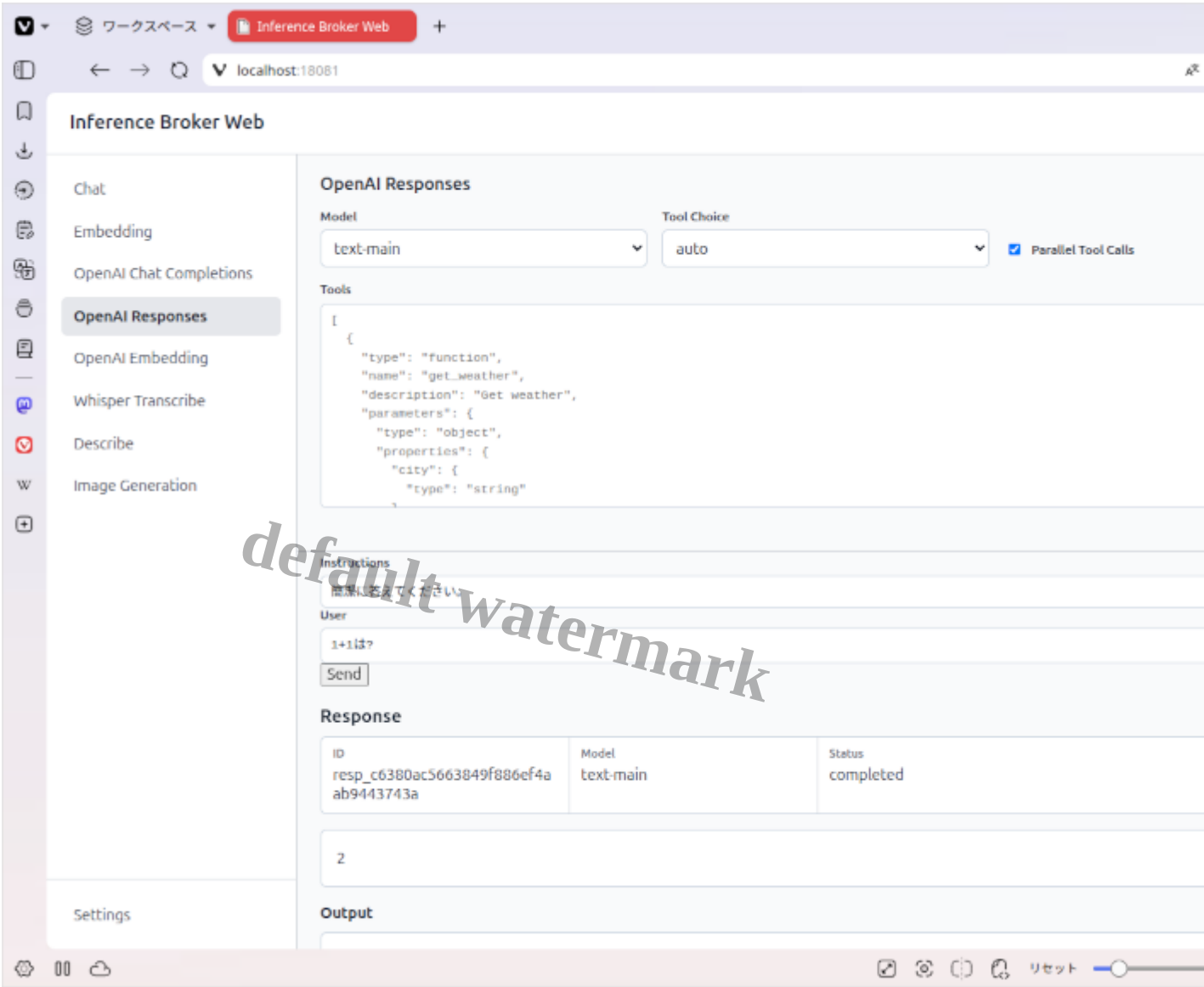
3. OpenAI /chat/completions

OpenAI API Chat API
「これはembeddingのテストです」



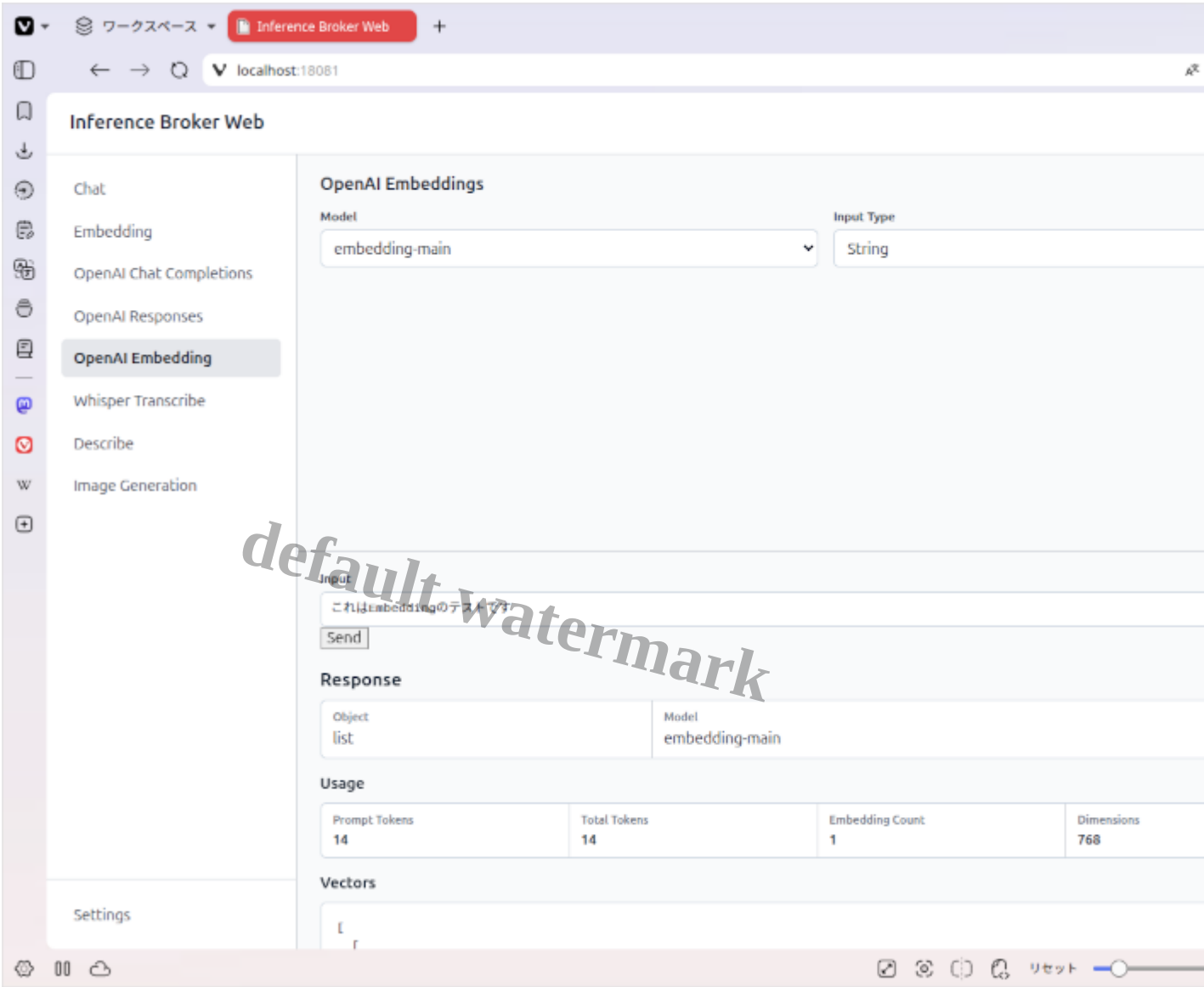
4. OpenAI Chat Responses

OpenAI Chat Responses API returns a JSON object with the following fields:



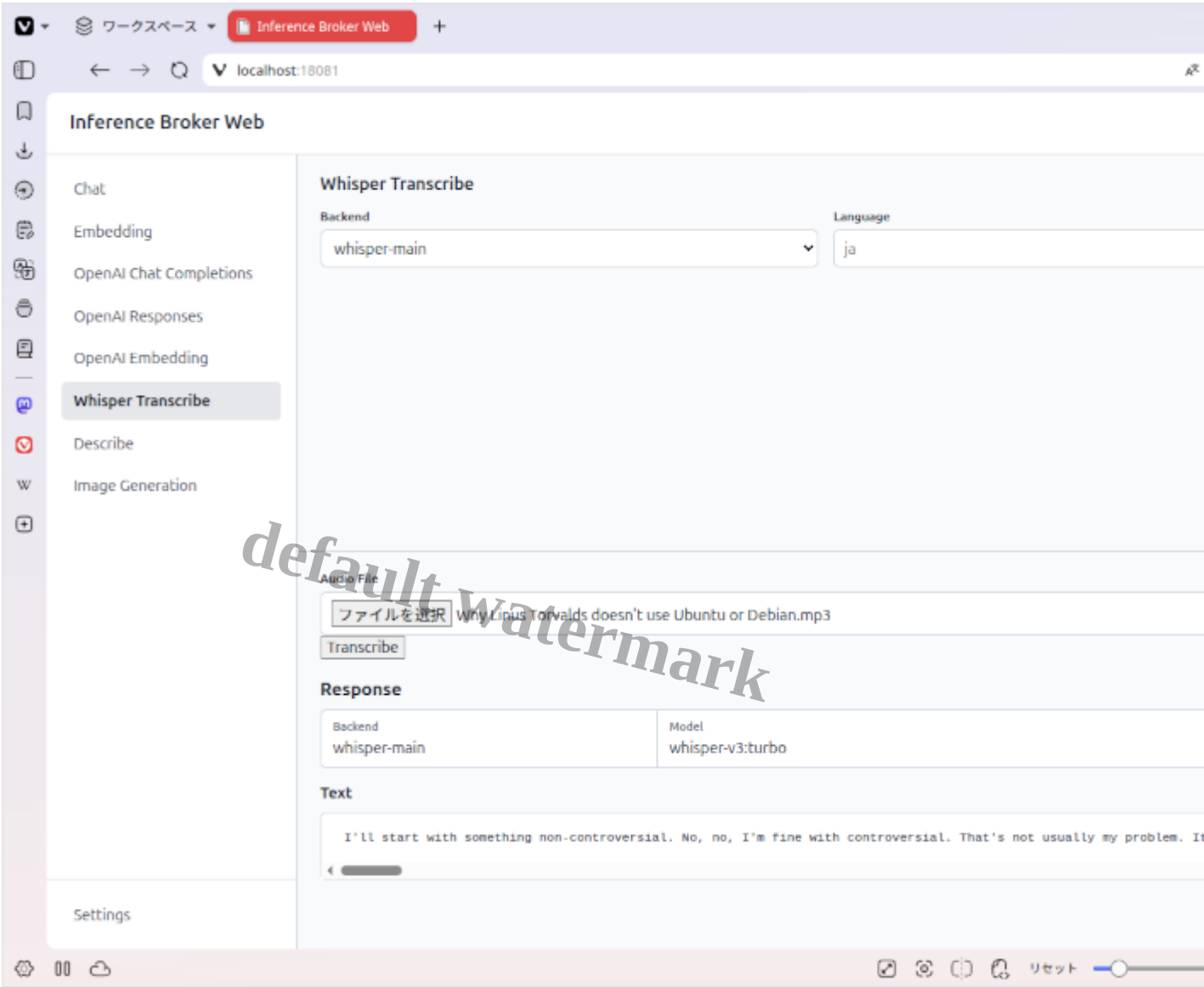
5. OpenAI Embeddings

OpenAI APIの使い方を紹介します。



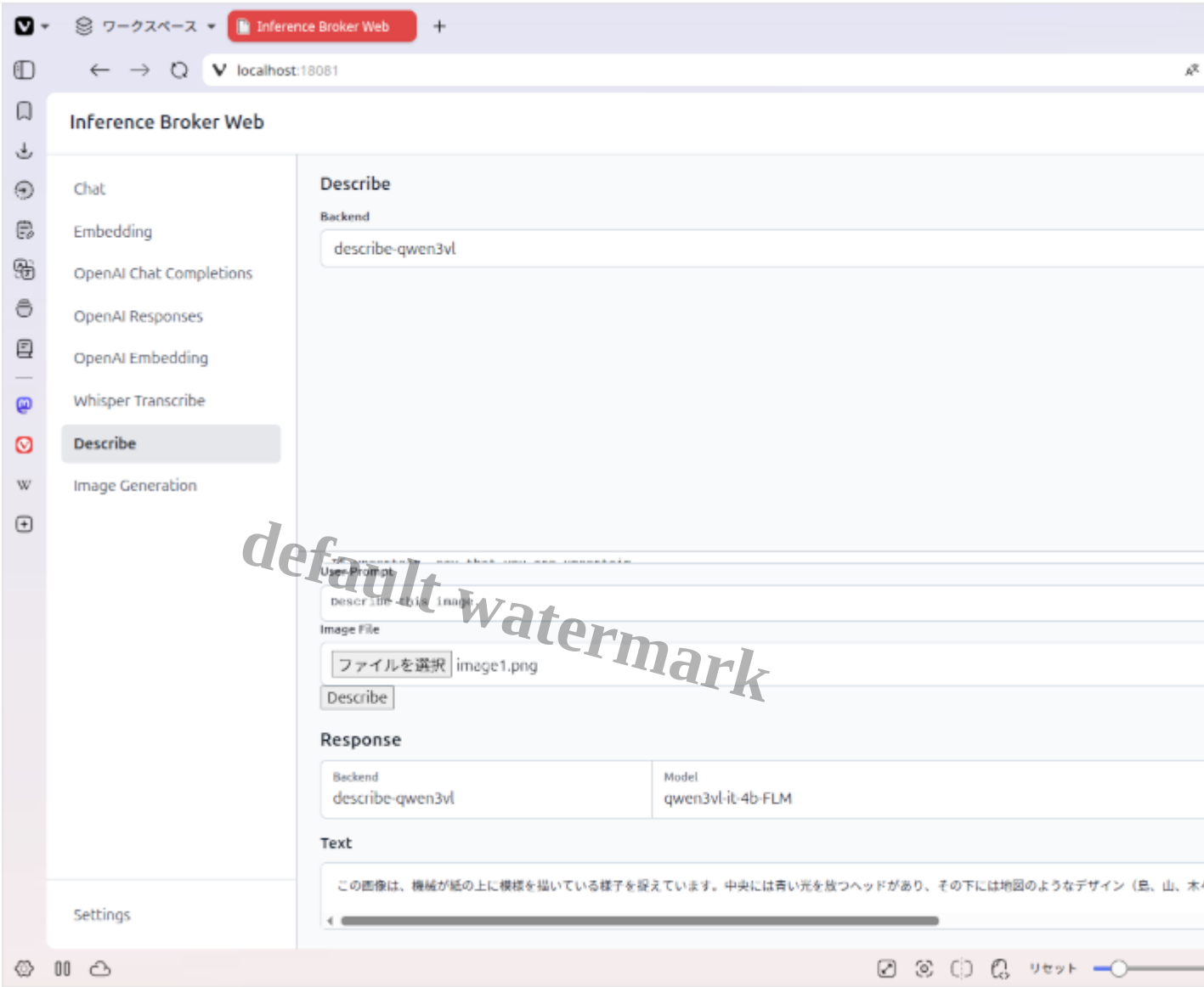
6. Whisper

NPUã??ã½¿ã•£ã•?ã?•é?³å£°ã??ã?jã?¤ã?«ã•?ã??æ??å?èµ.ã•?ã•?ã•?ã??ã•?ã•?ã•?®APIã•\$ã•?ã??



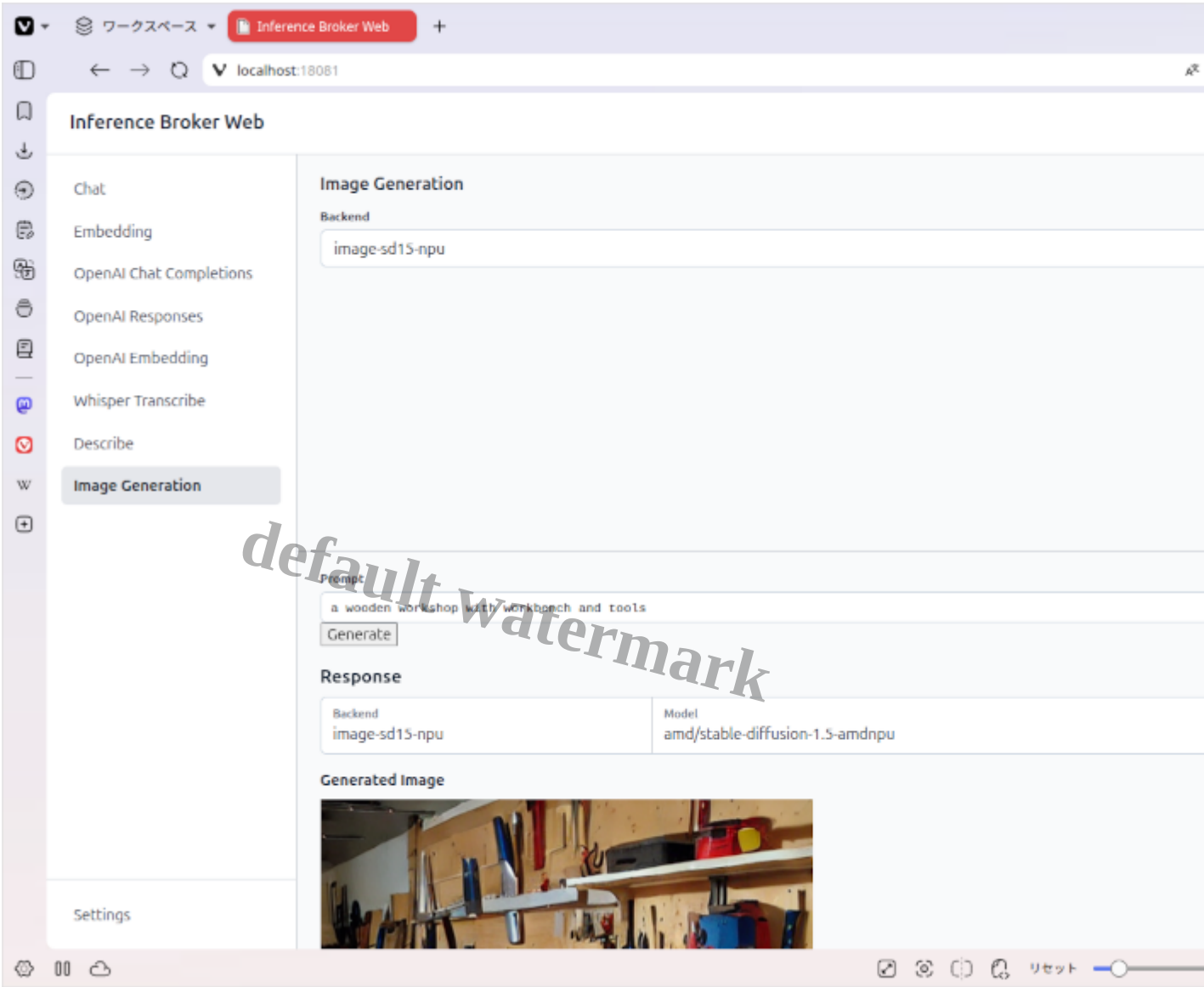
7. Describe

NPUä,?ã•\$Visionã?¢ã??ã?«ã??ã½¿ã•£ã•|ç?»ã?•ã??è\$£èª—ã•?ã??APIã•\$ã•?ã??OCRæ©?è?½ã??ã•
?ã??ã•¾ã•?ã??Qwen3 VL 4Bã??ã½¿ã•£ã•|ã•?ã•¾ã•?ã??

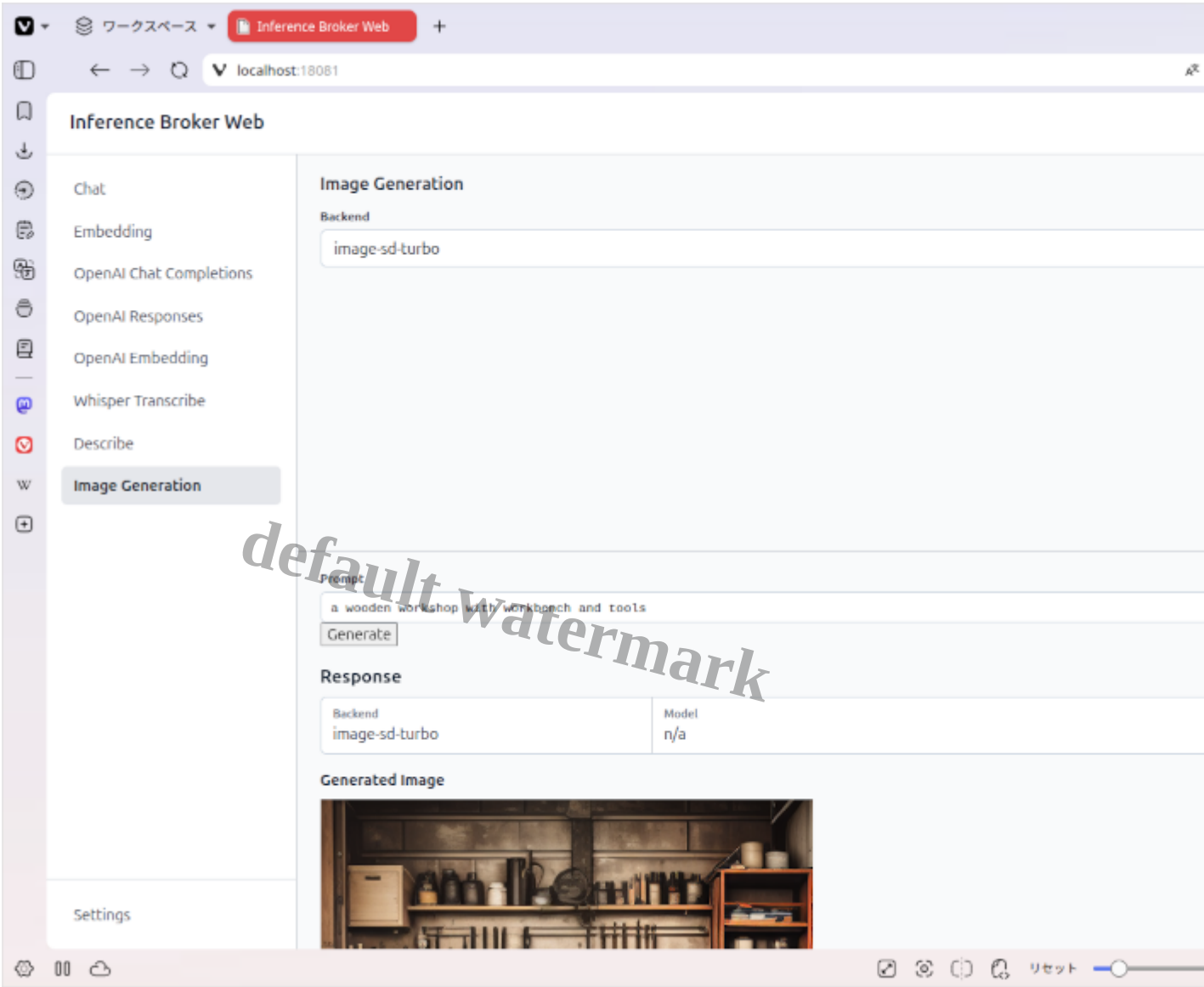


8. Image Generation

Stable Diffusionç³»ã•®ã?¢ã??ã?«ã??ã½¿ã•£ã•'ç?»å?•ç??æ?•ã??ã?ã??ã?ã?ã?ã•®APIã•§ã?ã??
NPUã??ã½¿ã•?ã??ã?ã•®ã•SD 1.5ç³»ã?¢ã??ã?«ã??ã½¿ã?ã?ã¼ã•?ã??



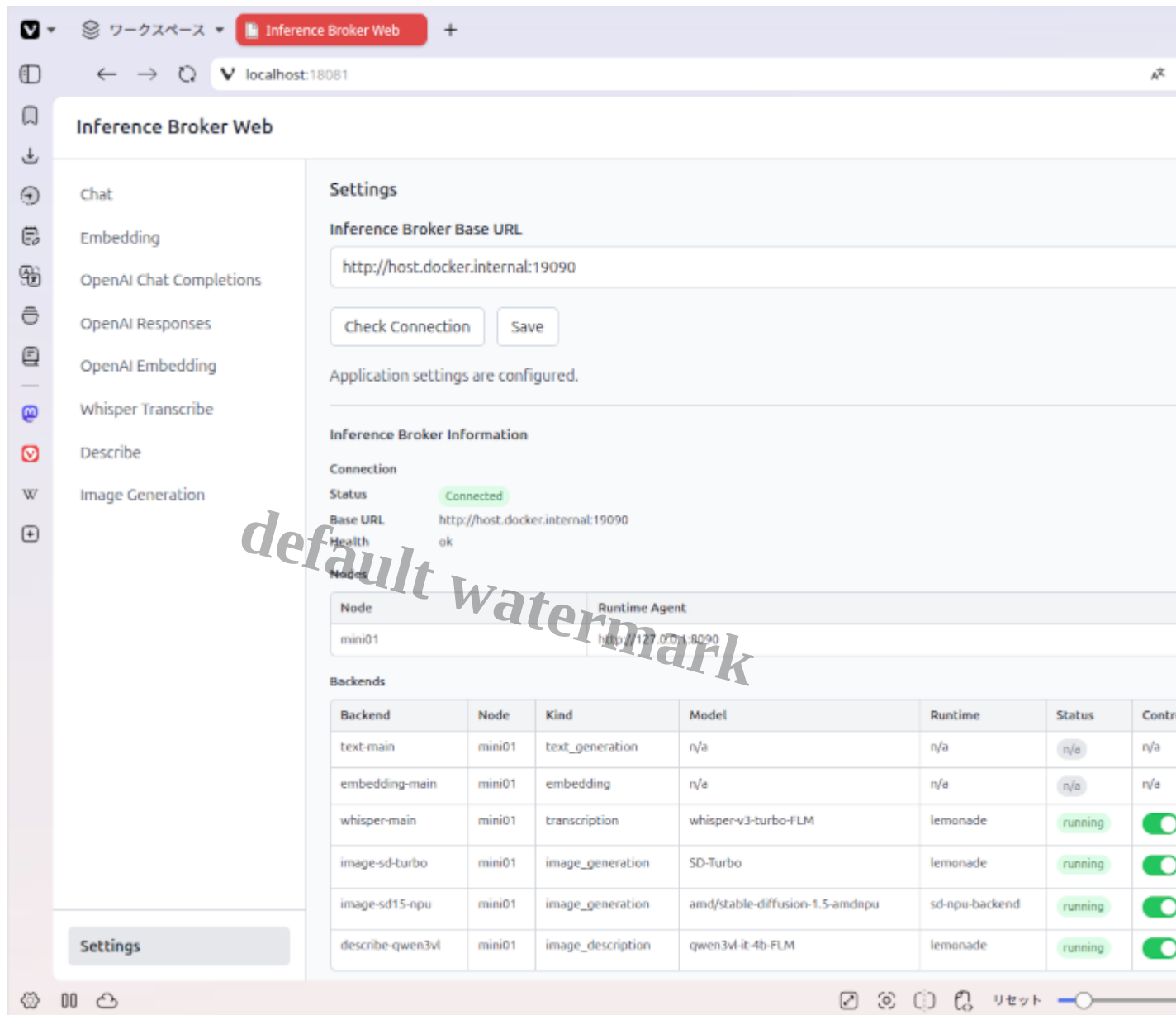
NPU??ǎ½¿ǎ?•ǎ•ǎ?ǎ??ǎ•Ⓢ̄SD Turboǎ??ǎ½¿ǎ?ǎ¾ǎ?ǎ??



Settings

Inference Broker Web

Inference Broker



ã?ã??ã?ã?ã?ã?ã?,?è|šã??è|ã??ã?ã?•Lemonadeã•šã??ã?•Whisper/SD-Turbo/Qwen3vlã?•SD NPU Backendã•ã?ã?ã?ZIKUUã•šã½?ã•£ã?•NPUã•šStable Diffusionã??ã??ã?ã?ã?ã?•®ã?ã?,!ã??ã•šã?ã?ã?ã?ã??

é•©å??ã•ªURLã??å ¥å??ã•?ã??ã•°ã•?Inference Brokerã•«æ?¥ç¶¶?ã•?ã??ã•!ã•?ã•,?è``ã•®8ç``®é¡?ã•
®APIã??è©!ã•?ã•¾ã•?ã??

[illegible]

ã??ã??ã?¹ã???ç³»ã•«ã•▯ã•?ã•|ã•-ã?¿LLM Runtime Agentã•®ã?¿¶ã¾jã-¾è±jã▯?ãªã•®ã•§ã?•ã•?ã•?ã??ã
•èµ.ã??ã?»ã?æ£ã•§ã•ã¾ã•?ã??ã??

- â•®ã?¿?ã•?ã«ã?•2â•°ã•®ã?µã?¼ã?•ã?¼ã•\$3ã•ã•®æ??æ?,ç??æ?•ç³»æ?“è«?è³?æ°•¼?llama.cppã•\$ã•
 ?ã•£ã•?ã?Ollamaã•\$ã•?ã•£ã•?ã?¿i¼?ã???ç?“æ?•ã•?ã??ã•ã?•ã?¤ã??ãªã?±ã?¼ã?•ã?§ã?³ã•
 _text_generationã??ã½¿ã?ã•ã??ã?ã?ã?©ã??ã?³ã?ã?ã?ã?ã?ã•ã?ã•\$ã?•Inference Brokerã?•æ??é•©ã•
 ºæ?“è«?è³?æ°•ã?¿é,ã?¿ã•\$ã?|ç?•çµæ??ã?¿è¿?ã?ã•|ã•ã?¿ã•¾ã?ã??
 æ®µé??ç??ãªè|_æ“jã•®æ?jã•\$ã?»ç,®ã°ã?ã?è?•é??ã®³ã”¾ã¿?ã?ã•\$ã•ã??ã?ã?ã•ã«ã•ã??ã•¾ã?ã??

default watermark

- ## Tags

- ## Date Created

2026å¹¸æ??29æ?¥

Author

kazuotsubaki