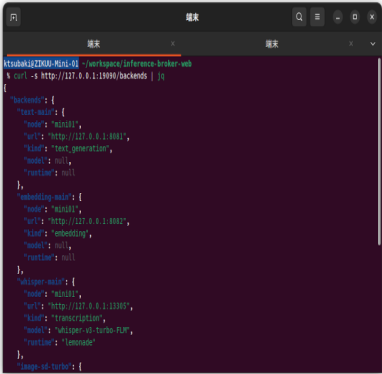




Inference Broker

Description

LLM Broker Vision Broker

LLM Runtime Agent SD NPU Backend Docker Sllama.cpp

LLM Runtime Agent

```
终端
ktsubaki@ZIKUU-Mini-01 ~/workspace/inference-broker-web
% systemctl status llm-runtime-agent.service
● llm-runtime-agent.service - LLM Runtime Agent
   Loaded: loaded (/etc/systemd/system/llm-runtime-agent.service; enabled; preset: enabled)
   Active: active (running) since Thu 2026-08-27 18:53:17 JST; 12h ago
     Main PID: 8820 (uvicorn)
        Tasks: 7 (limit: 68339)
      Memory: 32.0M (peak: 36.2M)
         CPU: 42.397s
    CGroup: /system.slice/llm-runtime-agent.service
            └─8820 /srv/llm-runtime-agent/.venv/bin/python3 /srv/llm-runtime-agent/.venv/bin/uvicorn app.map

8月 28 06:35:47 ZIKUU-Mini-01 uvicorn[8820]: INFO: 127.0.0.1:46408 - "GET /backends/sd-npu-backend/status HTTP/1.1" 200 OK
8月 28 06:40:11 ZIKUU-Mini-01 uvicorn[8820]: INFO: 127.0.0.1:41796 - "GET /backends/lemonade/status HTTP/1.1" 200 OK
8月 28 06:40:11 ZIKUU-Mini-01 uvicorn[8820]: INFO: 127.0.0.1:41804 - "GET /backends/lemonade/status HTTP/1.1" 200 OK
8月 28 06:40:11 ZIKUU-Mini-01 uvicorn[8820]: INFO: 127.0.0.1:41808 - "GET /backends/sd-npu-backend/status HTTP/1.1" 200 OK
8月 28 06:55:32 ZIKUU-Mini-01 uvicorn[8820]: INFO: 127.0.0.1:50418 - "GET /backends/lemonade/status HTTP/1.1" 200 OK
8月 28 06:55:32 ZIKUU-Mini-01 uvicorn[8820]: INFO: 127.0.0.1:50428 - "GET /backends/lemonade/status HTTP/1.1" 200 OK
8月 28 06:55:32 ZIKUU-Mini-01 uvicorn[8820]: INFO: 127.0.0.1:50430 - "GET /backends/sd-npu-backend/status HTTP/1.1" 200 OK
8月 28 07:03:00 ZIKUU-Mini-01 uvicorn[8820]: INFO: 127.0.0.1:36548 - "GET /backends/lemonade/status HTTP/1.1" 200 OK
8月 28 07:03:00 ZIKUU-Mini-01 uvicorn[8820]: INFO: 127.0.0.1:36550 - "GET /backends/lemonade/status HTTP/1.1" 200 OK
8月 28 07:03:00 ZIKUU-Mini-01 uvicorn[8820]: INFO: 127.0.0.1:36556 - "GET /backends/sd-npu-backend/status HTTP/1.1" 200 OK
lines 1-20/20 (END)
```

SD NPU Backend Ryzen AI CPU NPU
Whisper Vision

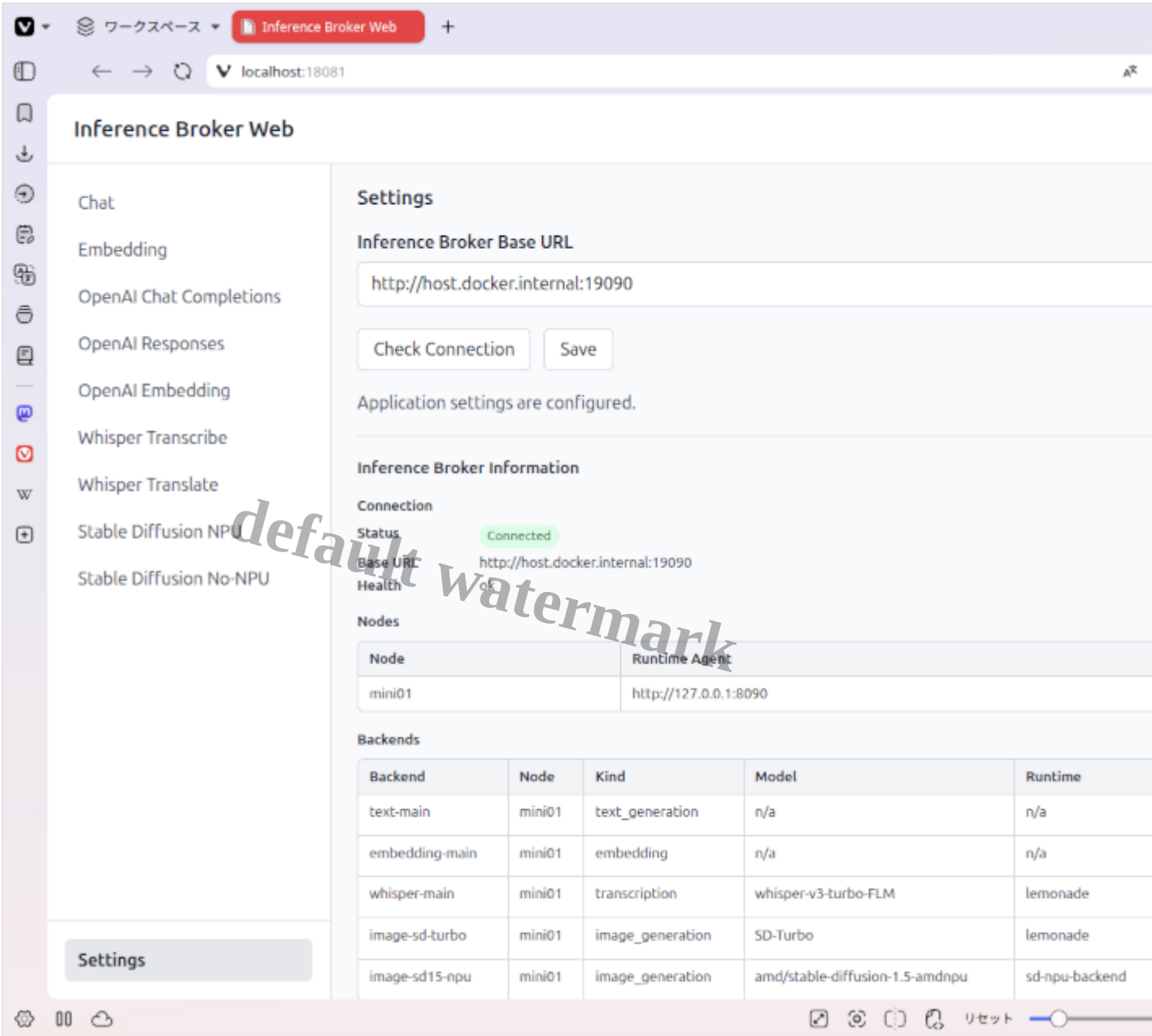
```
终端
终端
ktsubaki@ZIKUU-Mini-01 ~/workspace/inference-broker-web
% systemctl status sd-npu-backend --no-pager
● sd-npu-backend.service - SD NPU Backend
   Loaded: loaded (/etc/systemd/system/sd-npu-backend.service; enabled; preset: enabled)
   Active: active (running) since Thu 2026-08-27 21:33:28 JST; 9h ago
     Main PID: 34881 (uvicorn)
        Tasks: 113 (limit: 68339)
      Memory: 2.4G (peak: 2.4G)
         CPU: 5min 45.791s
    CGroup: /system.slice/sd-npu-backend.service
            └─34881 /srv/sd-npu-backend/.venv/bin/python /srv/sd-npu-backend/.venv/bin/uvicorn backend.app:..

8月 27 21:33:41 ZIKUU-Mini-01 bash[34881]: Git branch: HEAD
8月 27 21:33:41 ZIKUU-Mini-01 bash[34881]: PID: 34881
8月 27 21:33:41 ZIKUU-Mini-01 bash[34881]: UID: 1000
8月 27 21:33:41 ZIKUU-Mini-01 bash[34881]: [Thu Aug 27 12:33:41 2026 GMT]
8月 27 21:33:41 ZIKUU-Mini-01 bash[34881]: HOST: ZIKUU-Mini-01
8月 27 21:33:41 ZIKUU-Mini-01 bash[34881]: EXE: /usr/bin/python3.12
8月 27 21:33:41 ZIKUU-Mini-01 bash[34881]: [XRT] WARNING: Legacy TXN flow is deprecated. Please migr-n/ELF.
8月 27 21:33:44 ZIKUU-Mini-01 bash[34881]: [1.2K blob data]
8月 27 21:33:45 ZIKUU-Mini-01 bash[34881]: ---Running VAE Decoder
8月 27 21:33:45 ZIKUU-Mini-01 bash[34881]: INFO: 127.0.0.1:53046 - "POST /v1/images/generations -200 OK
Hint: Some lines were ellipsized, use -l to show in full.
ktsubaki@ZIKUU-Mini-01 ~/workspace/inference-broker-web
%
```

Inference Brokerã¬ã?ã??ã??ã??ã?©ç?¬ã?ã?ã?•æ??ç« ç??æ?ã?•ç?»ã?•ç??æ?•ã?•æ??ã?èµ·ã?ã?ã?ã?•ç?»ã?•è\$£æ?•ã?ã?©ã??è|?ã?æ?¬è«?è³?æ°•ã??ç®|ç?ã?ã?ã?£ã??ã?ªã?±ã?¼ã?·ã?§ã³ã?ã?ã??ã?®æ?¬è«?è|æ±?ã??æ?¬è«?è³?æ°•ã?«æ?¬ã??ã??ã??ã?½¹ã?²ã??æ?•ã?jã?¾ã?ã??

Inference Broker Webã¬ã?Inference Brokerã¬®æ©?è?½ã??æ¬?è¼¼ã?ã??ã?ã?ã?ã?ã?ã?®Webã?£ã??ã?ªã?±ã?¼ã?·ã?§ã³ã?ã?§ã?ã??

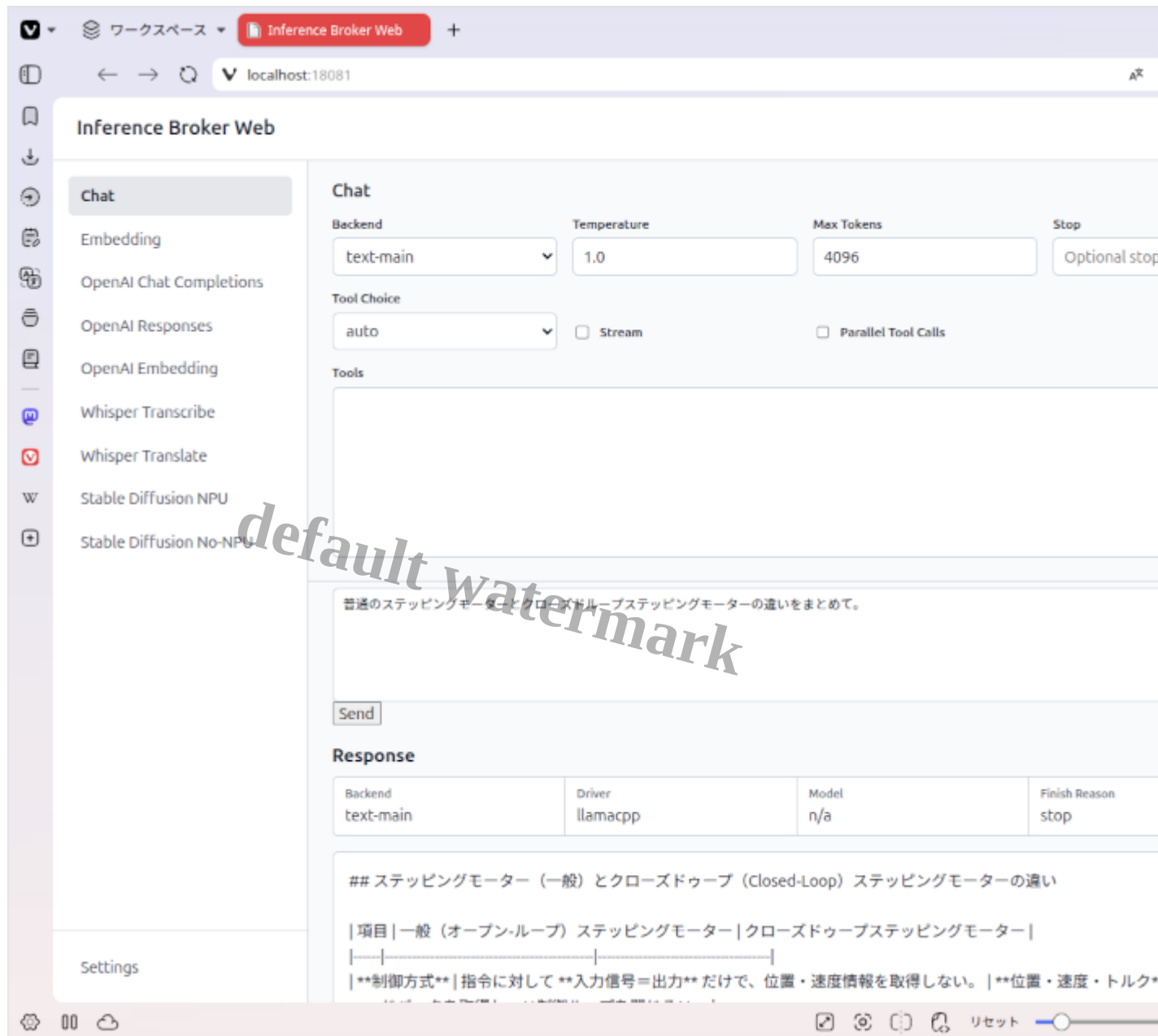
Inference Broker Webã¬®è"å®?ç?»é•£ã?§ã¬ã?ã??ã?¼ã??æ? å ±ã??ã?•ã??ã?¬ã?¬ã?³ã??æ? å ±ã?è|?ã?ã?¾ã?ã??



ã•?ã•®ä,?ã•§ã?•

- Chat
- Embedding
- OpenAlä°?æ•? Chat Completions
- OpenAlä°?æ•? Embedding
- OpenAlä°?æ•? Responses
- Whisper Transcribe
- Whisper Translate
- Stable Diffusionã•®NPUç??ã•é??ã„ç??

ã•®APIä??ã??ã?¹ã??ã•§ã•ã??ã??ã?ã•«ã•ã??ã•¼ã•?ã??



ä,?ã•ã®Inference Brokerã??é??ã•?ã|ã?•è•?æ?°ã•®ã??ã?¹ã??i¼?ã??ã?¼ã??i¼?ã?Sèµ°ã??LLM Runtime Agentã•?çµ±å?¶ã•?ã??ã•|ã?ã•ã•©ã•?ã•Sã?»ã•©ã??ã•æ?“è«?è³?æ°•ã•?å??ã•?ã•|ã•?ã??ã•?ã??æ?•è?ã•?ã•?ã•«ã?ã•?ã?±ã?¼ã?ã?Sã?³ã•?æ?“è«?å?|ç?ã??è?ã•?ã??ã??ã•?ã•«ã•ã??ã•¼ã•?ã??ã•?ã??ã•?ZIKUUè£½ã•®æ?“è«?ã??ã?©ã??ã??ã??ã?©ã?¼ã?ã•ã•?ã•|æ?“æ°?ã•«ã•ã??ã•¼ã•?ã??

ç•¼å?“ã?•Inference Brokerã•ZIKUU Miniæ©?ã•Sé??ç?“ã•?ã•|ã•?ã•¼ã•?ã•?ã•?ã•®?æ?•ã•?ã•?ã??ZIKUUã??ã?»ã?ã?¹ã??ã?ã•«é•å??ã•?ã•¼ã•?ã??

Category

- æ?¥ã?ã•®æ´»å??

Tags

-
1. AI
 2. Inference Broker
 3. LLM
 4. LLM Broker
 5. LLM Runtime Agent
 6. Ryzen AI
 7. SD NPU Backend
 8. Stable Diffusion
 9. Vision Broker
 10. Whisper
 11. ZIKUU Mini
 12. ä,?é??å??å,?
 13. å±±æ¢“ç??

Date Created
2026å¹¸æ??28æ?¥

Author
kazuo-tsubaki

default watermark