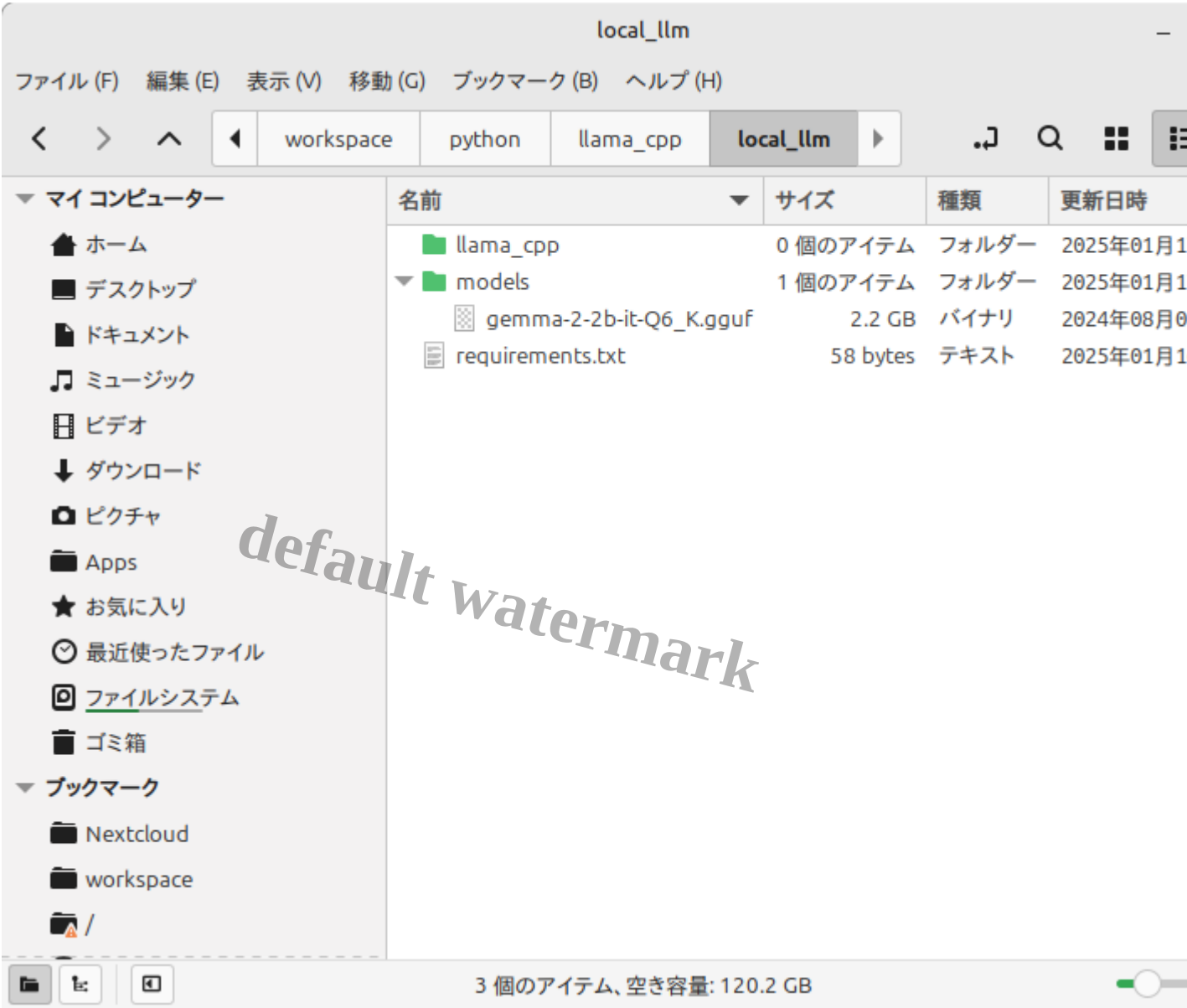




æ??ã??å?ç?ã•ã??ã?ã°ã?©ã?

Pythonã•Sè``èª?ã?¢ã??ã?«ã??ã½¿ã•£ã•|æ??ç« ç??æ?•ã??èj?ã?ç°jã?ã•ã??ã?ã?ã°ã?©ã?ã??ã½?ã•£ã•|ã?•æ£ã?ã•ã?|ç?ã?ã?ã•Sã•ã??ã?ã•ã??ç?¢èª•ã?ã•¾ã?ã??

main.pyã??æ¬jã•®å? å®¹ã•Sã½?ã??ã•¾ã?ã??

from llama_cpp import Llama

DEFAULT_SYSTEM_PROMPT = "ã•?ã•ªã•?ã•`èª å®?ã•Sã?ªçS?ã•ªæ?¥æ?-ã°ª?¢ã?•ã?¹ã?¿ã?³ã??ã•Sã•?ã??è³ªã?ã•«ã½ã?ã•|æ?¥æ?-èª?ã•Sã, •ã`Sã•«ã??ç?ã?ã•|ã•ã•ã?ã?ã?ã?"

```
def make_prompt(message):
    prompt = "<bos><start_of_turn>user {system} {prompt} <end_of_turn> <start_of_turn>assistant"
    return prompt
```

```
llm = Llama(
    model_path="./models/gemma-2-2b-it-Q6_K.gguf",
    n_ctx=4096,
)
output = llm(make_prompt("Name the planets in the solar system?. "),
             max_tokens=4096,
             echo=True
            )
print(output)
```

main.pyã??å®?è¡?ã•?ã•¾ã•?ã??

```
python main.py
```

æ-jǎ•Ⓜă??ă•?ă•^aă°ă°??ă•?ă³/₄?ă??ă??ă??ă°æŁă,,ă«ă??ă¹/₂?ă•?ă|ă•?ă³/₄ă•?ă??i¹/₄?ă^aé?¹/₂c³»ă•
Ⓜæ??æ??ă•Ⓜă•ă°ă•?ă??è^{3a}ă°ă•?ă•?ă°ă•?ă^aă•?ă•Śă•?ă•?ă•ă•ă??ă??ă•?ă•?ă°ă°ç?ă•?ă³/₄?ă??ă??ă•
?i¹/₄?

```
llama_model_loader: loaded meta data with 39 key-value pairs and 288 tensors f
llama_model_loader: Dumping metadata keys/values. Note: KV overrides do not ap
llama_model_loader: - kv 0:                          general.architecture str
llama_model_loader: - kv 1:                          general.type str
llama_model_loader: - kv 2:                          general.name str
llama_model_loader: - kv 3:                          general.finetune str
llama_model_loader: - kv 4:                          general.basename str
llama_model_loader: - kv 5:                          general.size_label str
llama_model_loader: - kv 6:                          general.license str
llama_model_loader: - kv 7:                          general.tags arr[s
llama_model_loader: - kv 8:                          gemma2.context_length u32
llama_model_loader: - kv 9:                          gemma2.embedding_length u32
```

```

.....
.....
.....
.....
llama_perf_context_print:      load time =      1265.40 ms
llama_perf_context_print: prompt eval time =          0.00 ms /      38 tokens (
llama_perf_context_print:      eval time =          0.00 ms /      93 runs (
llama_perf_context_print:      total time =    10987.63 ms /    131 tokens
{'id': 'cmpl-ea481b2d-bee9-4158-8b9b-77e435daee45', 'object': 'text_completion
•?ã•ªã•?ã•_èª å@?ã•šã?ªçš?ã•ªæ?¥æ?-ä°ã?çã?•ã?¹ã?¿ã?³ã??ã•šã•?ã??è³ªã?•ã•
«ã_¼ã•?ã•|æ?¥æ?-èª?ã•šã_•ã_šã•«ã??ç?ã•?ã•|ã••ã•_ã•?ã•
?ã?? Name the planets in the solar system?. <end_of_turn> <start_of_turn>mode
•«ã_ã?•ãªªé?½ã??ã_ã¿?ã•«ã??ã??æ??æ??ã•?ã?•ã»¥ã_?ã•@é??ã??ã•šã•
?ã??\n\n1. **Mercury:** é??æ??\n2. **Venus:** å??æ??\n3. **Earth:** å?°ç•
?\n4. **Mars:** ç•
«æ??\n5. **Jupiter:** æ?¨æ??\n6. **Saturn:** é??æ??\n7. **Uranus:** åª©ç??æ

```

ä•?ä•®ç?ä•?ä??ä?ä?°ä?©ä? ä•?è``èª?ä?¢ä??ä?«ä??æ?±ä•?ä??ä?ä?°ä?©ä? ä•®èµ·ç?¹ä•«äªä??ä•¼ä
 •?ä??

Web UI

Webã??ã?-ã¼ã? ã?¬ã¼ã?¬ã®StreamLitã??ã½¿ã•£ã!Webã??ã?©ã?|ã?¶ã•ã??ã?£ã?¬ã?»ã?¹ãšã
•ã??ã??ã?ã?°ã?©ã? ã??ã½?ã??ã•¾ã•?ã??

web.pyã•ã•?ã•?ã•?ã•?ã•®ã??ã?|ã?ã?«ã??ã½?ã??ã?•æ¬|ã•®ã? å®¹ã«ç·é??ã•?ã|ä¿•åã?ã•?ã¾ã?
?ã??

```
import streamlit as st
from langchain.callbacks.base import BaseCallbackManager
from langchain_community.llms import LlamaCpp
from langchain.callbacks.base import BaseCallbackHandler
```

```
DEFAULT_SYSTEM_PROMPT = "ã·?ã·ªã·?ã·`èª ¤®?ã·§ã?ªç§?ã·  
ªæ?¥æ?-äºã·¢ã··ã·¹ã·¿ã·³ã·??ã·§ã·?ã·?è³ª¤·?ã·«ã³ã·?ã·|æ?¥æ?-èª?ã·§ã··ã·§ã·  
«ã?¿ç?ã·?ã·|ã··ã·ã·?ã·?ã·?"
```

```
model_path = "./models/gemma-2-2b-it-Q6_K.gguf"
```

```
class StreamHandler(BaseCallbackHandler):
    def __init__(self, container, initial_text="", display_method="markdown"):
        self.container = container
        self.text = initial_text
        self.display_method = display_method

    def on_llm_new_token(self, token: str, **kwargs) -> None:
        self.text += token
        display_function = getattr(self.container, self.display_method, None)
        if display_function is not None:
            display_function(self.text)
        else:
            raise ValueError(f"Invalid display_method: {self.display_method}")
```

```
def make_prompt(message):
    prompt = "<start_of_turn>user {system} {prompt} <end_of_turn> <start_of_turn>assistant"

    return prompt
```

```
with st.sidebar:
    st.header('æ?¥æ?¬èª?ç??æ?•AI')
    st.subheader('ä½?ã??ã?»å!ã•¶ ä?¢ã??ã•¥ã••ã??å¡¼ã??ZIKUUã?•')
    st.divider()

    st.text("è`å@?")
    max_tokens = st.slider('Max Tokens', value=40960, min_value=8192, max_value=40960)
    temperature = st.slider('Temperature', value=0.8, min_value=0.1, max_value=1.0)
    n_ctx = st.slider("Number of Ctx", value=8192, min_value=64, max_value=8192)
    n_batch = st.slider("Number of Batch", value=128, min_value=1, max_value=8192)
```

```
with st.form(key="generation_form"):
    prompt = st.text_area('è³aâ?•')
    do_generate = st.form_submit_button('é?•ä¿i')
    if do_generate and prompt:
        with st.spinner("ç??æ?•ä..."):
            chat_box = st.empty()
```

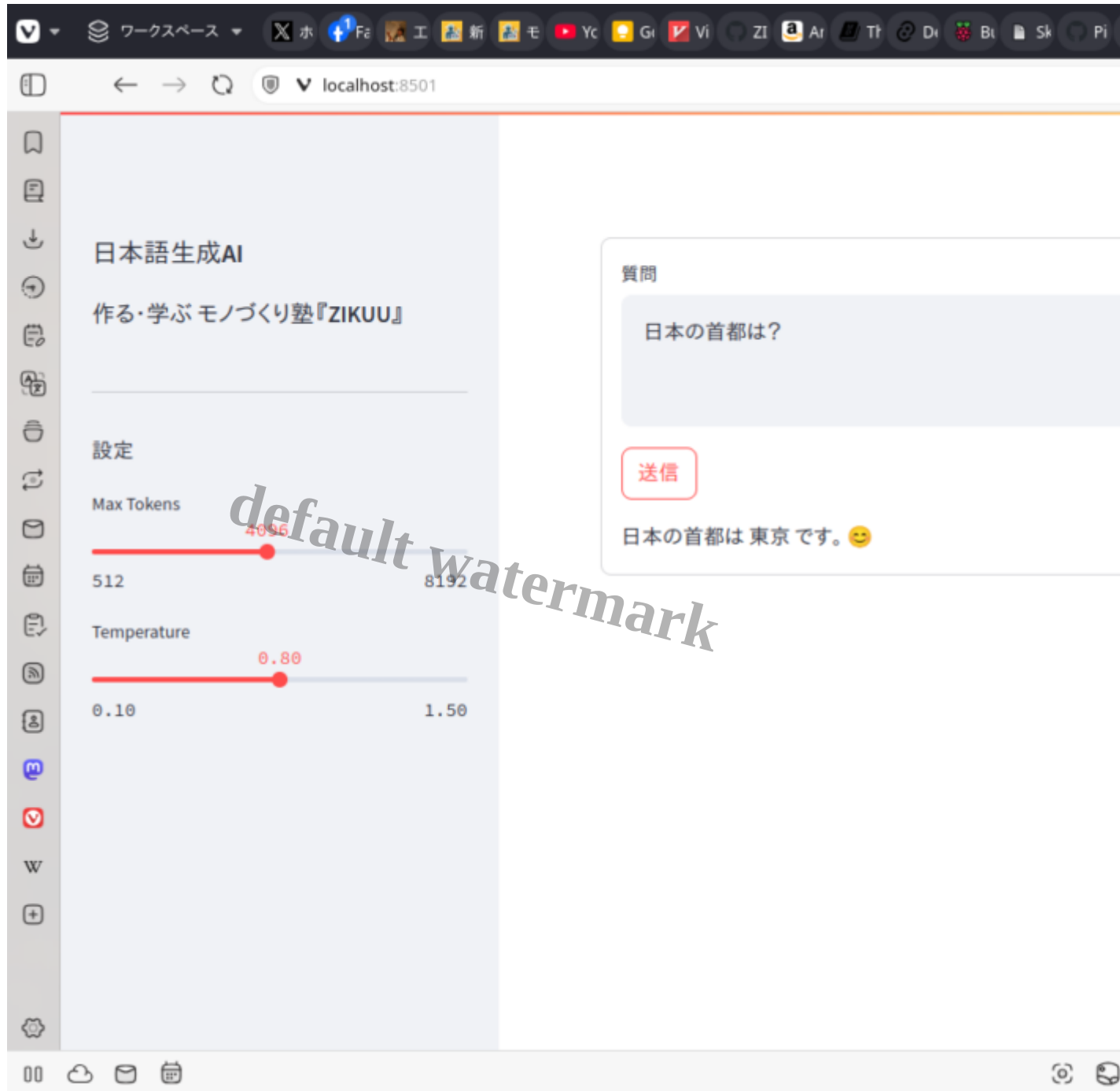
```
stream_handler = StreamHandler(chat_box, display_method='write')
chat = LlamaCpp(
    model_path=model_path,
    temperature=temperature,
    max_tokens=max_tokens,
    n_ctx=n_ctx,
    n_batch=n_batch,
    callback_manager=BaseCallbackManager([stream_handler]),
    verbose=False,
)
res = chat.invoke(make_prompt(prompt))
```

æ¬ã•®ã³ã??ã³ã??ã•šã??ã?ã°ã?©ã?ã??èµ·å??ã?ã¼ã•?ã??

streamlit run web.py

Webã??ã?©ã?|ã?¶ã•?ã?? http://localhost:8501ã•«ã?¢ã?¬ã?»ã¹ã?ã??ã•ã?æ¬ã•®ã??ã•?ã•ç?»é•¢ã•
?èj`çºã•?ã??ã?•è³ªã?•æ¬ã•«ã½?ã•?ã??å¥ã??ã•?ã•é?•ä¿ã??ã?¿ã³ã??æ?¼ã•?ã•ç??æ?•ã•?ã??ã•
?æ??ç«ã•?èj`çºã•?ã??ã??ã•°æ?•å??ã•šã•?ã??

default watermark



ã?ã??ã?šç??æ?•Alã??æ?±ã?•Webã?¢ã??ã?ªã?±ã?¼ã?·ã?šã?³ã??ã½?ã??ã?ã?ã?ç?®ç??ã¾ã?šé?
?æ?•ã?šã?ã¾ã?ã?ã??

Dockerã?šã??ã?ã?°ã?©ã?ã??ã??ã?ã?ã?

Dockerã?OSã®ã,?ã«ã»®æ?³ã®OSç?°å¢?ã??ã½?ã??ã?¢ã??ã?ªã?±ã?¼ã?·ã?šã?³ã??é?ç?ã?
?ã??ã?ã?ã?ã®ã??ã?¼ã?«ã?šã?ã??ã½?æ?ã?ã?ã?ã?¢ã??ã?ªã?±ã?¼ã?·ã?šã?³ã??Dockerã?šã??ã?ã?
?ã??ã??ã?ã?«ã?
«ã?¢ã??ã?ªã?±ã?¼ã?·ã?šã?³ã??ã?
ã?,ã?šã?ã??ã??ã?ã?±æ??ã?ã??ã?

Dockerfile
FROM ubuntu:22.04
ENV DEBIAN_FRONTEND=noninteractive
ENV DEBCONF_NOWARNINGS=yes
RUN apt-get update && apt-get upgrade -y
RUN apt-get dist-upgrade -y && apt-get autoremove -y
RUN apt-get install -y build-essential wget
RUN apt-get install -y python3 python3-venv
WORKDIR /app
RUN wget -P ./models https://huggingface.co/bartowski/gemma-2-2b-it-GGUF/resolve/main/gemma-2-2b-it-Q4_K_M.gguf?download=true
COPY ./requirements.txt /app
COPY ./web.py /app
RUN python3 -m venv /llama_cpp
RUN . /llama_cpp/bin/activate && python -m pip install --upgrade pip
RUN . /llama_cpp/bin/activate && python -m pip install -r requirements.txt
COPY ./docker/entrypoint.sh /
RUN chmod 755 /entrypoint.sh
EXPOSE 8501
ENTRYPOINT ["/entrypoint.sh"]

mkdir docker
cd docker

Dockerfile

```
1: FROM ubuntu:22.04
2:
3: ENV DEBIAN_FRONTEND=noninteractive
4: ENV DEBCONF_NOWARNINGS=yes
5:
6: RUN apt-get update && apt-get upgrade -y
7: RUN apt-get dist-upgrade -y && apt-get autoremove -y
8: RUN apt-get install -y build-essential wget
9: RUN apt-get install -y python3 python3-venv
10:
11: WORKDIR /app
12:
13: RUN wget -P ./models https://huggingface.co/bartowski/gemma-2-2b-it-GGUF/resolve/main/gemma-2-2b-it-Q4_K_M.gguf?download=true
14:
15: COPY ./requirements.txt /app
16: COPY ./web.py /app
17:
18: RUN python3 -m venv /llama_cpp
19: RUN . /llama_cpp/bin/activate && python -m pip install --upgrade pip
20: RUN . /llama_cpp/bin/activate && python -m pip install -r requirements.txt
21:
22: COPY ./docker/entrypoint.sh /
23: RUN chmod 755 /entrypoint.sh
24:
25: EXPOSE 8501
26: ENTRYPOINT ["/entrypoint.sh"]
```

FROM ubuntu:22.04

- 1. FROM ubuntu:22.04
- 2. ENV DEBIAN_FRONTEND=noninteractive
- 3. ENV DEBCONF_NOWARNINGS=yes
- 4. RUN apt-get update && apt-get upgrade -y
- 5. RUN apt-get dist-upgrade -y && apt-get autoremove -y
- 6. RUN apt-get install -y build-essential wget
- 7. RUN apt-get install -y python3 python3-venv
- 8. WORKDIR /app
- 9. RUN wget -P ./models https://huggingface.co/bartowski/gemma-2-2b-it-GGUF/resolve/main/gemma-2-2b-it-Q4_K_M.gguf?download=true
- 10. COPY ./requirements.txt /app
- 11. COPY ./web.py /app
- 12. RUN python3 -m venv /llama_cpp
- 13. RUN . /llama_cpp/bin/activate && python -m pip install --upgrade pip
- 14. RUN . /llama_cpp/bin/activate && python -m pip install -r requirements.txt
- 15. COPY ./docker/entrypoint.sh /
- 16. RUN chmod 755 /entrypoint.sh
- 17. EXPOSE 8501
- 18. ENTRYPOINT ["/entrypoint.sh"]

- 22èj?ç?@i¼?entrypointã??ã?jã?ðã?«i¼?ã??ã?ã?°ã?©ã? ã•@å@?èj?é??å§?ã?³ã??ã?³ã??i¼?ã•
@ã?³ã??ã?¼
- 23èj?ç?@i¼?entrypointã??ã?jã?ðã?«ã,ã•@å@?èj?æ"©ã•@ä»?ä,?
- 25èj?ç?@i¼?ã?¢ã?¬ã?»ã?¹ã?ã??ã?•ã??ã??ã?¬ã?¼ã?¬ã?¼ã??ã?@å -é??i¼?Streamlitã•
@ã??ã??ã?©ã?«ã??ã?•ã?¼ã??ã?§ã?i¼?
- 26èj?ç?@i¼?entrypointã•@æ??å@?

æ¬jã«docker-compose.ymlã"ã•?ã?å•å?ã•@Dockerã??ã??ã?ã?ðã??ã?jã?ðã?«i¼?Docker
composeã??ã?jã?ðã?«i¼?ã??ã?½?ã??ã?¾ã?ã??

version: '3.8'

services:

app:

build:

context: .

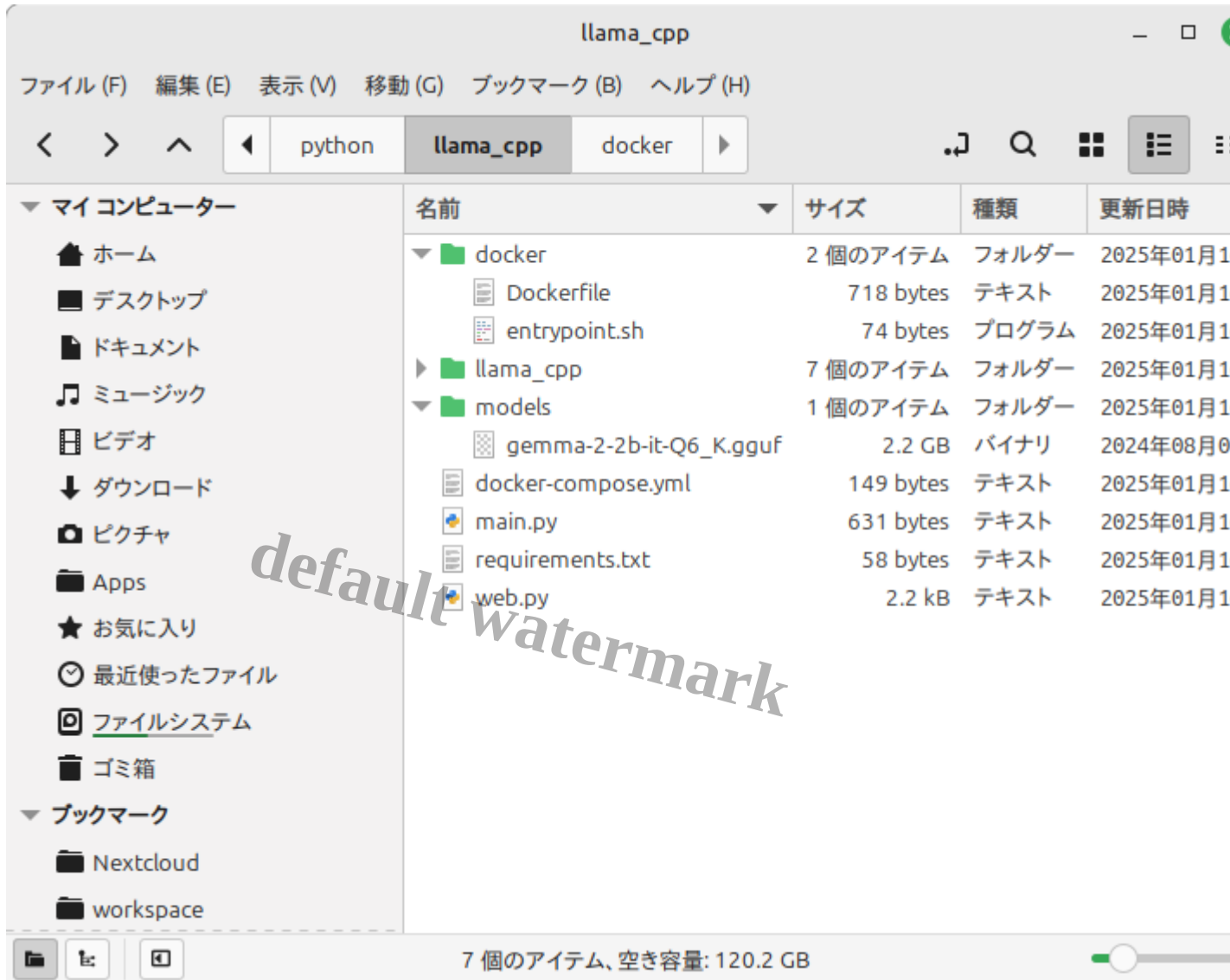
dockerfile: ./docker/Dockerfile

network: host

ports:

- "8501:8501"

ã•?ã•@æ@mé??ã•§ã??ã?ã?,ã?§ã?¬ã??ã??ã?©ã?«ã??ã?¼ã¬æ¬jã@ã??ã?ã•«ã•ã•£ã!ã•?ã??ç?ã•§ã•
?ã??



ã•?ã??ã•\$Dockerã•\$ã?¢ã??ã?ªã?±ã?¼ã?·ã?§ã?³ã??å??ã?•ã?•æº?å??ã?•æ?´ã?•ã?¾ã?•ã?•ã??

æ-jǎ•®ǎ?³ǎ??ǎ?³ǎ??ǎ•sǎ?¢ǎ??ǎ?ªǎ?±ǎ?¼ǎ?·ǎ?sǎ?³ǎ??èµ.ǎ??ǎ?·ǎ?¾ǎ?·ǎ??i¼?i¼?ǎ??ç?®ǎ•
®èµ.ǎ??æ??ǎ•è¨?èª?ǎ?¢ǎ??ǎ?«ǎ??ǎ?©ǎ?¤ǎ??ǎ?©ǎ?ªǎ®ǎ?¤ǎ?³ǎ?¹ǎ??ǎ?¼ǎ?«ǎ•«æ??é??ǎ?·ǎ?·ǎ•
?ǎ??ǎ?¾ǎ?·i¼?

```
docker compose up
```

Webã??ã?©ã?|ã?¶ã?¼ã?š http://localhost:8501 ã??é??ã?•ã?•ã?•ã? ?ç??ã?•ã?•ã?•ã?»é•£ã?•?è|ç?ªã?•
?ã??ã??ã?•æ?•ã??ã?šã?•ã??

ä•?ä•®Dockerä??ä½¿ä•?æ?¹æ³?ä•ä??ä•°ä?•ä?µä?¼ä?•ä?¼ä•«Pythonä?©ä?¤ä??ä?©ä?ªä•ä•©ä??ä•
 ?ä•jä•?ä•jä?¤ä?³ä?¹ä??ä?¼ä?«ä•?ä•ªä•ä•
 ¡ä??ä?¢ä??ä?ªä?±ä?¼ä?•ä?§ä?³ä??ä?¤ä?³ä?¹ä??ä?¼ä?«ä?»å®?èj?ä•?ä•§ä•ä•¾ä?ä•?ä•?
 é??ç?°ç?“ä•®PCä•ä?µä?¼ä?•ä?¼ä•ä•®OSä•®é?ä•?ä??æ?èèä?ä•ä•ä•ä!æ?ä•¿ä•¾ä?ä??

å®?é???ã•®é??ç?°ã?»é•?ç?`ã•®ç?¾å´ã•šã¯ã?ã??ã?,ã?šã¯ã??ã??Githubã??Gitlabã•ã•©ã•®ã?ªã?
ã?ã?ã?ã?ªã«ç?»é²ã•?ã!ã?ã•ã?ã?ãµã¹¼ã•ã?¼ã•šã?ªã?ã?ã??ã?ªã•?ã??ã??ã?-

ã?_ã?šã?_ã??ã??ã??ã?|ã?³ã?ã?¼ã??ã?ã?|ã?•docker compose upã?³ã??ã?³ã??ã?•
šã?šã??ã?ªã?±ã?¼ã?_ã?šã?³ã??èµ_ã??ã?ã??ã?_ã?ã?ã?ã?•æ?¹æ³?ã??ã?ã??ã?¼ã?_ã??ã?ã?šè|_æ¨_jã?
ªã?_ã?¹ã??ã?_ã?šã?_ã?µã?¼ã?_ã?¼ã?_ã?_ã?_ã?©ã?¹ã?_ã?¼ã?ã??ã?ã??ã??ã?ªã?©ã?®è|_ç´_ã?ã?ã?_ã?ã??ã?
¾ã?_ã?ã?ã?_ã?°æ?_ç??ã?ªé??ç?°ã?_ã??ã?é?ç?°ã?¾ã?šã?®æ??é_ã?_ã?ã?ã?ã?ã?ã?ã?®ã??ã?ã?ã?¼šã?šã?
•?ã??i¼?CI/CD â?? Continuous Integration/Continuous Development = ç¶¶?ç¶¶?ç??ã?ªé?ç?°ã?
¨é??ç?°ã?_ã?_ã?¼ã?_ã??ã??ã??ã?ã?®ã?šã?_i¼?

ãµšè|_æ¨_jè¨?èªã?šã??ã?«ã??ã?½ç_ã?ã?_ç?°jã?ã?ã?ã?šã??ã?±ã?¼ã?_ã?šã?³ã??é??ç?°ã?ã?ã?ã?ã?ã?
?ã?_ã?ã?ã?æµ_ã??ã??èššèª_ã?_ã?¾ã?_ã?
ã?¥ã?¼ã?ã??ã?
šè¨?èªã?šã??ã?«ã??é??ç?°ã?ã??ã?ã?ã?ã?»ã?®ã??ã?ã?ã?ã?ã?ã?ã?ã?ã?ã?ã?ã?ã?ã?ã?ã?ã?ã?ã?
šã?šã??ã?ªã?±ã?¼ã?_ã?šã?³ã??ã?®?èš_ã?ã??ã?_ã?
©ã?_ã?ã?¶ã?ã??ã?¼ã?ã??ã?

ä»?»ã??ã?®ã?_ã?®¹ã?_ç??ã?¼¼ã?ã??ã?ã?ã?ã?ã?°æ?¥ã?_ã?ã?ã?æ?°ã?ã?ã?ã?¾ã?ã??ã?ã?i¼?

æ??ã?_ã?_ç??ã?¼¼ã?šè¨_ã?ã?®ã?šã?ã??ç??ã?¼¼ã?šã??è?_ã?ã?ã?ã?æ??ã??ã?ã?ã?ã?ã?ã?ã?ã?ã?ã?ã?
?ã?ã?«ã??ã?³ã??æ??ã?ã?è??ã?³ã?_æ?±ã?¾ã??ã?ã?ã?ã?®ã?_ã?ã?«é?²ã?_ã?è?_ã?ã?®ã?šã?ã??

Category

- 1. æ?¥ã?_ã?®æ´_»ã??
- 2. ç?_ã??è¨?

Tags

- 1. AI
- 2. Docker
- 3. Linux
- 4. LLM
- 5. Python
- 6. ä_?é??ã??ã?_?
- 7. ãµšè|_æ¨_jè¨?èªã?šã??ã?ã?«
- 8. ã±±æç¨ç??

Date Created

2025ã¹¹1æ??17æ?¥

Author

kazuo-tsubaki