

Web??ā?-ā?¼ā? ā?-ā?¼ā?-ā?@StreamLit??ā½¿ā•Ēā•Web??ā?©ā?!ā?¶ā•?ā??ā?ā?ā?-ā?»ā?¹ā•šā
•ā??ā??ā?ā?°ā?©ā? ā??ā½?ā??ā•¾ā•?ā??

web.pyā•ā•?ā•?ā•ā•?ā•@ā??ā?jā?ā?«ā??ā½?ā??ā?•æ-īā•@ā? ā®¹ā•«ç.é??ā•?ā•ä¿•ā?ā•?ā•¾ā•
?ā??

```
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 = "ā•?ā•ā•?ā•-èª ā@?ā•šā?ªçš?ā•  
ªæ?¥æ?-ā°ā?çā?•ā?¹ā?¿ā?³ā??ā•šā•?ā??è³ªā?•ā•«ā¾ā•?ā•|æ?¥æ?-èª?ā•šā,•ā-šā•  
«ā??ç?ā•?ā•|ā••ā• ā•?ā•?ā??"
```

```
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('ā½?ā??ā?»ā|ā•¶ ā?çā??ā•¥ā••ā??ā;¾ā??ZIKUUā?•')
    st.divider()
```

```
st.text("è"ā@?")
max_tokens = st.slider('Max Tokens', value=40960, min_value=8192, max_value=81920)
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('è³ªā?•')
    do_generate = st.form_submit_button('é?•ä¿j')
    if do_generate and prompt:
        with st.spinner("ç??æ?•ā,..."):
            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-ıã•@ã?³ã??ã?³ã??ã•\$ã??ã?ã?°ã?@ã? ã??èµ.â??ã?ã?¼ã?ã??

streamlit run web.py

Webã??ã?©ã?ıã?¶ã?ã?? http://localhost:8501ã?«ã?çã?ã?»ã?'ã?ã??ã?ã?æ-ıã•@ã??ã?ã?ã?ç?»é•çã•
?èj"çªã?ã??ã?•è³ã?ã?æ-ıã?ã?«ã?½?ã?ã??ã?¥ã??ã?ã?é?•ä¿ã??ã?¿ã?³ã??æ?¼ã?ã?ç??æ?ã?ã??ã?
?æ??ç« ã?èj"çªã?ã??ã??ã?°æ?ã??ã?\$ã?ã??

default watermark

Dockerfile
FROM ubuntu:22.04

ENV DEBIAN_FRONTEND=noninteractive
ENV DEBCONF_NOWARNINGS=yes

```
mkdir docker  
cd docker
```

Dockerfile
FROM ubuntu:22.04

```
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/r  
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
- 6. FROM ubuntu:22.04
- 13. FROM ubuntu:22.04
- 18. FROM ubuntu:22.04
- 20. FROM ubuntu:22.04
- 25. FROM ubuntu:22.04

