


```
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.py

```
python main.py
```

```
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)
```

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'  
Name the planets in the solar system?. <end_of_turn> <start_of_turn>mode  
**Mercury:**  
**Venus:**  
**Earth:**  
**Mars:**  
**Jupiter:**  
**Saturn:**  
**Uranus:**
```

Web UI

Web??ā?-ā?¼ā? ā?-ā?¼ā?-ā?@StreamLit??ā½¿ā•Ēā•Web??ā?©ā?!ā?¶ā•?ā??ā?çā?-ā?»ā?¹ā•§ā
•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•ªā•?ā•-èª ā@?ā•§ā?ªç§?ā•
ªæ?¥æ?-ā°ª?çā?•ā?¹ā?¿ā?³ā??ā•§ā•?ā??è³ªā?•ā•«ā¾ā•?ā•|æ?¥æ?-èª?ā•§ā,•ā-§ā•
«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_tu

    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_valu
    temperature = st.slider('Temperature', value=0.8, min_value=0.1, max_value
    n_ctx = st.slider("Number of Ctx", value=8192, min_value=64, max_value=819
    n_batch = st.slider("Number of Batch", value=128, min_value=1, max_value=8
```

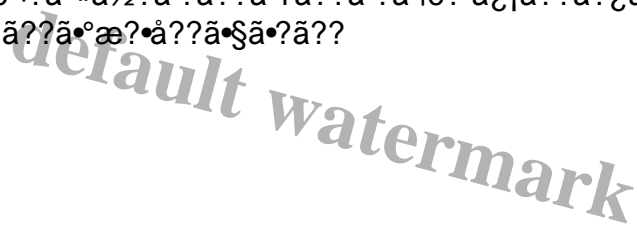
```
with st.form(key="generation_form"):
    prompt = st.text_area('è³ªā?•')
    do_generate = st.form_submit_button('é?•ä¿j')
    if do_generate and prompt:
        with st.spinner("ç??æ?•ā,..."):
            chat_box = st.empty()
```

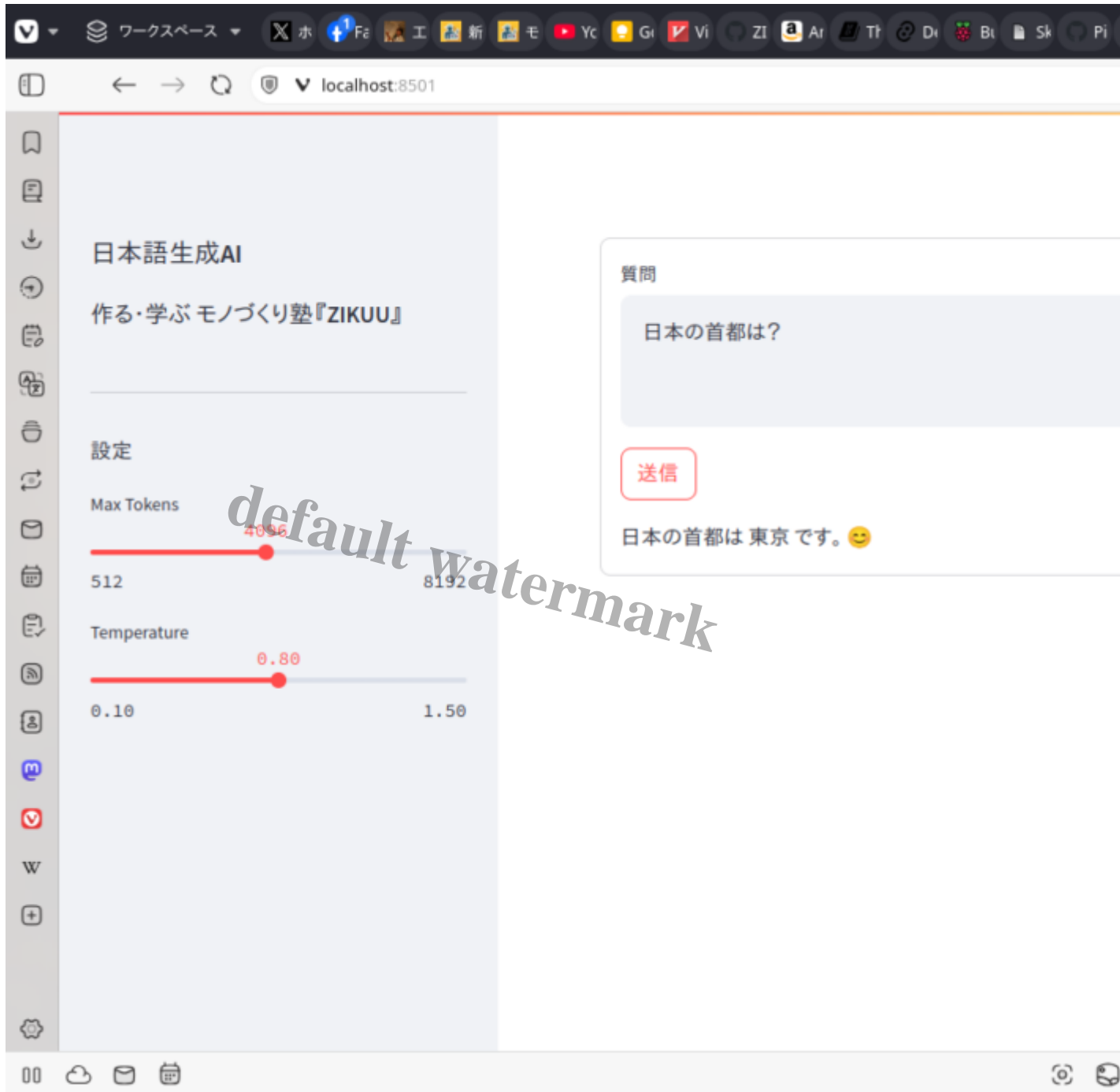
```
stream_handler = StreamHandler(chat_box, display_method='write')
chat = LlamaCpp(
    model_path=model_path,
    temperature=temperature,
    max_tokens=max_tokens,
    n_ctx=n_ctx,
    n_batch=n_batch,
    callback_manager=BaseCallbackManager([stream_handler]),
    verbose=False,
)
res = chat.invoke(make_prompt(prompt))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

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?ç?»é•çã•
?èj"çªã?ã??ã?•è³ã?ã?æ-ıã?ã?«ã?½?ã?ã??ã?¥ã??ã?ã?é?•ã?ıã??ã?ıã?³ã??æ?¼ã?ã?ç??æ?ã?ã??ã?
?æ??ç« ã?èj"çªã?ã??ã??ã?°æ?ã??ã?\$ã?ã??





ã?ã??ã?Şç??æ?•Alã??æ?±ã?•Webã?çã??ã?ªã?±ã?¼ã?·ã?Şã?³ã??ã½?ã??ã?ã?ã?ç?®ç??ã³¼ã?Şé?
?æ?ã?Şã?ã³¼ã?ã?ã??

Dockerã?Şã??ã?ã?°ã?©ã?ã??ã??ã?ã?ã?

Dockerã?OSã?®ã?,?ã?«ã?»®æ?³ã?®OSç?°ã?ç?ã??ã½?ã??ã?çã??ã?ªã?±ã?¼ã?·ã?Şã?³ã??é?ç?"ã?
?ã??ã?ã?ã?ã?®ã??ã?¼ã?«ã?Şã?ã??ã½?æ?•ã?ã?ã?ã?çã??ã?ªã?±ã?¼ã?·ã?Şã?³ã??ã?ã?
?ã??ã??ã?ã?ã?«ã?ã?ã?ã?ã?ã?ã?ã?µã?¼ã?ã?¼ã?
«ã?çã??ã?ªã?±ã?¼ã?·ã?Şã?³ã??ã?ªã?³ã?¹ã??ã?¼ã?«ã?ã?ã??ã?ã?»?ã?®ã?ª?ã??ã?-
ã?,ã?Şã?ã??ã??ã?±æ??ã?ã??ã?®ã?ã?®æ??ã?Şã?ã??

Dockerfile

FROM ubuntu:22.04

```
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/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" ]
```

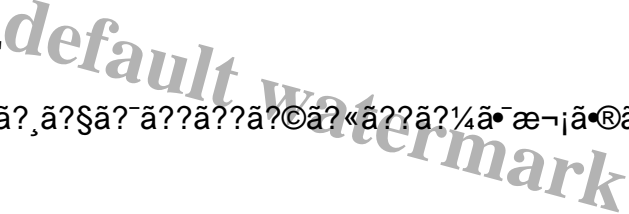
FROM ubuntu:22.04

- 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

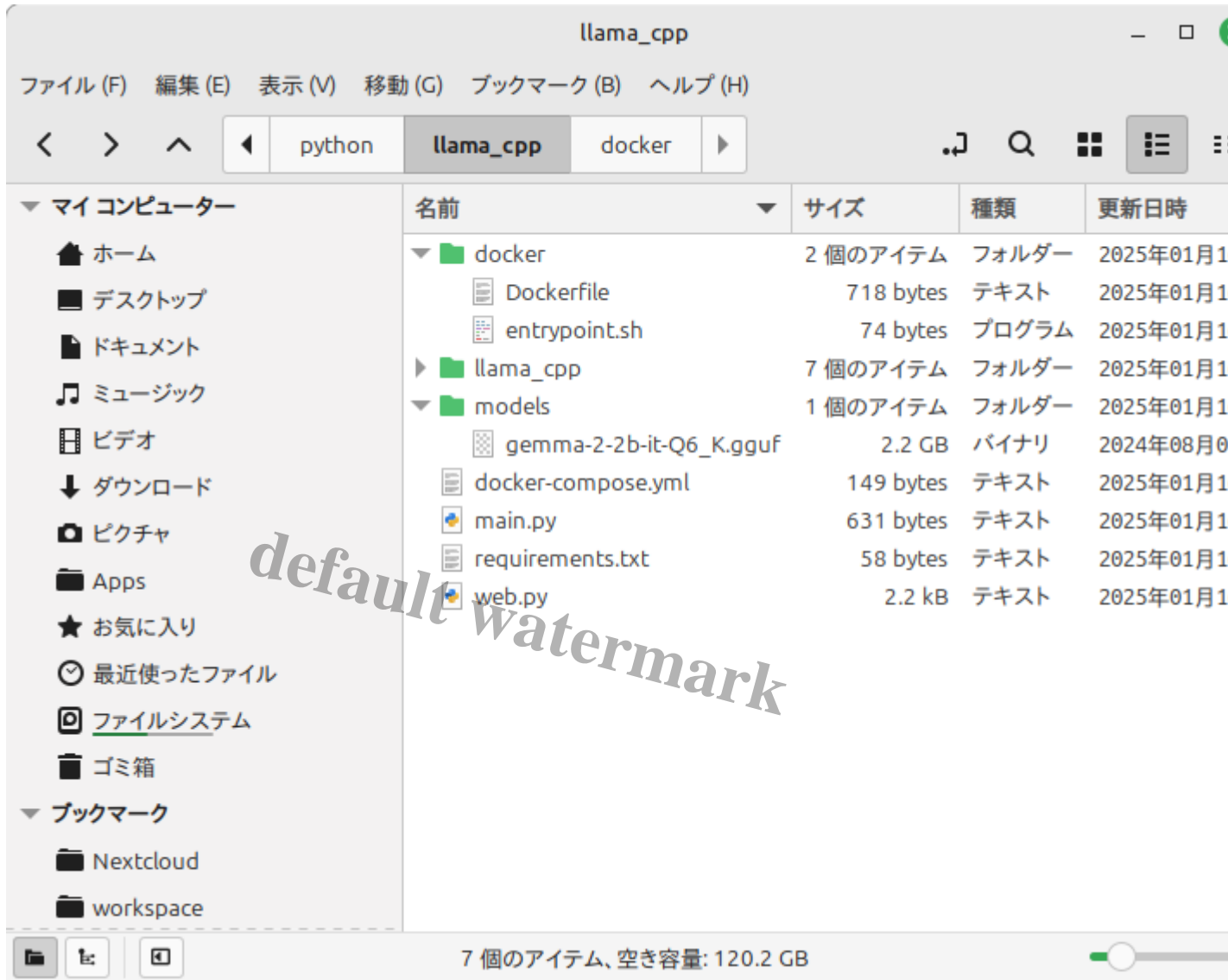
- 22 è j ? ç ? @ i ¼ ? entrypoint ä ? ? ä ? j ä ? ð ä ? « i ¼ ? ä ? ? ä ? ä ? ° ä ? © ä ? ä • @ ä @ ? è ; ? é ? ? å § ? ä ? ³ ä ? ? ä ? ³ ä ? ? i ¼ ? ä • @ ä ? ³ ä ? ? ä ? ¼
- 23 è j ? ç ? @ i ¼ ? entrypoint ä ? ? ä ? j ä ? ð ä ? « ä , ä • @ ä @ ? è ; ? æ " © ä • @ ä » ? ä , ?
- 25 è j ? ç ? @ i ¼ ? ä ? ç ä ? - ä ? » ä ? ¹ ä • ? ä ? ? ä ? • ä ? ? ä ? ? ä ? - ä ? ¼ ä ? - ä ? • ä ? ¼ ä ? ? ä • @ ä - é ? ? i ¼ ? Streamlit ä • @ ä ? ? ä ? ? ä ? © ä ? « ä ? ? ä ? • ä ? ¼ ä ? ? ä • § ä • ? i ¼ ?
- 26 è j ? ç ? @ i ¼ ? entrypoint ä • @ æ ? ? å @ ?

æ - j ä • « docker-compose.yml ä " ä • ? ä • ? å • å ? • ä • @ Docker ä ? ? ä ? ? ä ? ä ? ð ä ? ? ä ? j ä ? ð ä ? « i ¼ ? Docker compose ä ? ? ä ? j ä ? ð ä ? « i ¼ ? ä ? ? ä ½ ? ä ? ? ä • ¾ ä • ? ä ? ?

```
version: '3.8'
services:
  app:
    build:
      context: .
      dockerfile: ./docker/Dockerfile
      network: host
    ports:
      - "8501:8501"
```



ä • ? ä • @ æ @ mé ? ? ä • § ä ? ? ä ? ä , ä ? § ä ? - ä ? ? ä ? ? ä ? © ä ? « ä ? ? ä ? ¼ ä • æ - j ä • @ ä ? ? ä • ? ä • « ä • ä • £ ä • ! ä • ? ä ? ? ç ? ä • § ä • ? ä ? ?



ã•?ã??ã•\$Dockerã•\$ã?çã??ã?ã?±ã?¼ã?-ã?§ã?³ã??ã??ã•?ã•?æ°?ã??ã•?æ?´ã?ã•¼ã•?ã•?ã??

æ-1ã•®ã?³ã??ã?³ã??ã•\$ã?çã??ã?ã?±ã?¼ã?-ã?§ã?³ã??èµ-ã??ã•?ã•¼ã•?ã??¼?i¼?ã??ç?®ã•
®èµ-ã??æ??ã?è"èªã?çã??ã?«ã??ã?®ã?ªã??ã?®ã?ªã?®ã?ªã?ªã?ªã?ªã?¼ã?«ã•æ??é??ã•?ã•ã?
?ã??ã•¼ã•?i¼?

docker compose up

Webã??ã?®ã?iã?¶ã?¼ã•\$ http://localhost:8501 ã??é??ã•ã"ã?ã?ã? ?ç"ã"ã?ã•?ç?»é•çã•?èj"çªã•
?ã??ã??ã°æ?ã?ã?ã•\$ã•?ã??

ã•?ã•®Dockerã??ã½¿ã•?æ?¹æ³?ã•ã??ã°ã?ã?ã?µã?¼ã?ã?¼ã•«Pythonã?®ã?ªã??ã?®ã?ªã?ã?ã?ã?
?ã•jã?ã?jã?ªã?³ã?¹ã??ã?¼ã?«ã•?ãªã•ã•
iã??ã?çã??ã?ã?±ã?¼ã?-ã?§ã?³ã??ã?ªã?³ã?¹ã??ã?¼ã?«ã?»ã®?èj?ã?ã•\$ã•ã¼ã•?ã•?ã?
é??ç?°ç?"ã•®PCã"ã?µã?¼ã?ã?¼ã"ã•®OSã•®é?ã?ã??æ?è?ã•?ãªã•ã•æ¿ã?ã½ã•¼ã?ã??

ã®?é??ã•®é??ç?°ã?»é?ç?"ã•®ç?¼ã ´ã•\$ã-ã?ã??ã?ã?,ã?§ã?-ã??ã??Githubã??Gitlabãªã•®ã•®ãªã?
ã?ã?ã??ãªã«ç?»é?²ã?ãªã•ã?ã?µã?¼ã?ã?¼ã•\$ã?ªã?ã?,ã??ã?ªã?ã??ã??ã?-

