

CTransformersã•Gradioã•?ã½¿ã•£ã•?Code Llamaã•?ã•?ã•?ã•?
?ã?£ã•?ã•?ã•?±ã•?¼ã•?ã•?Sã•?³ã•??Dockerã•?¾ã•?¿ã•?ã•?¾ã•?ã•?ã•?ã•?ã•?£ã•?ã•?¥ã•?ã•?ã•?¿ã•?¾ã•?®ã•?µã•?¼ã•?ã•?
ã•?¼ã•?Sã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?¾ã•?ã•?ã•?

The screenshot shows the VS Code editor with the following content:

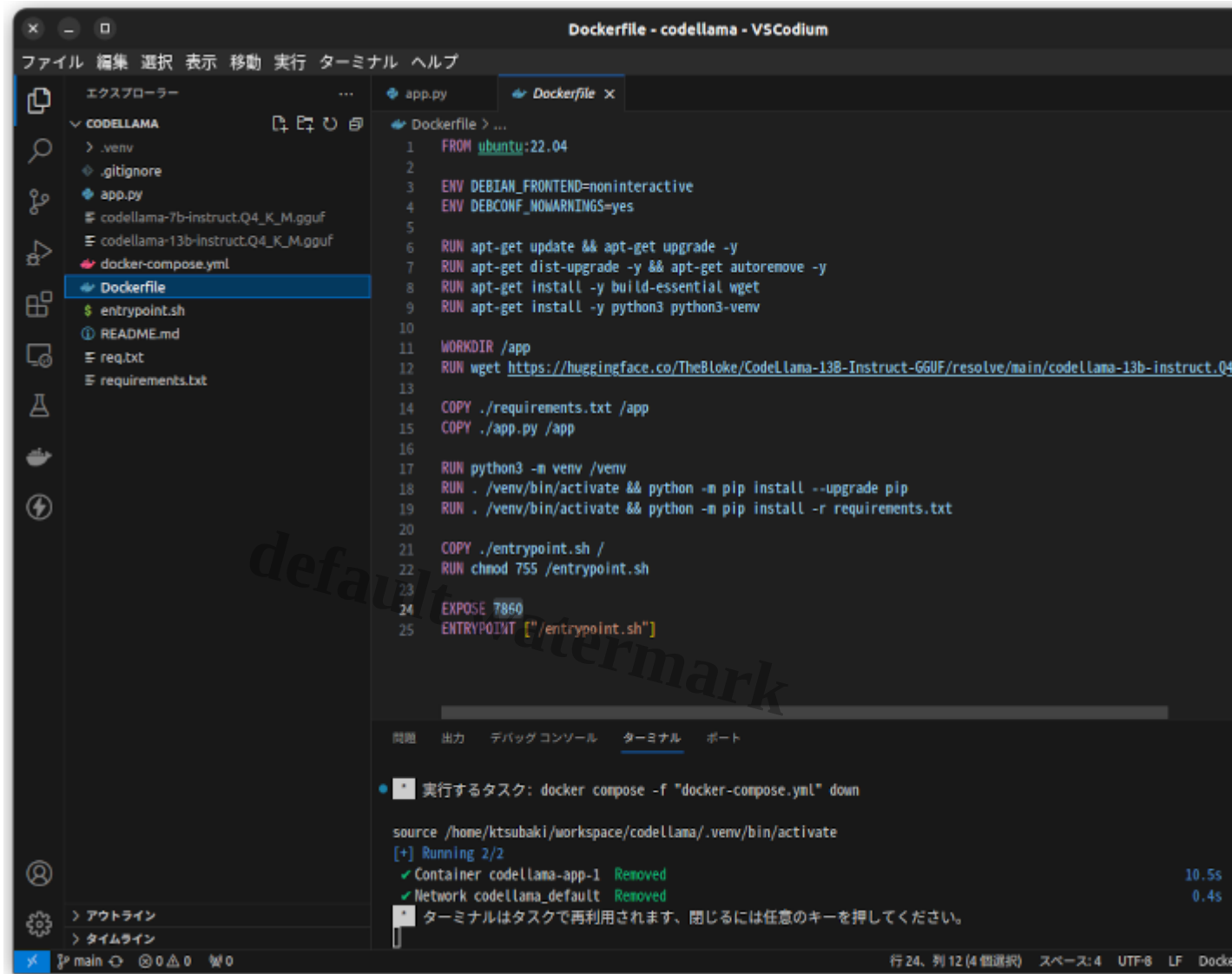
```
app.py - codellama - VSCodeium

ファイル 編集 選択 表示 移動 実行 ターミナル ヘルプ

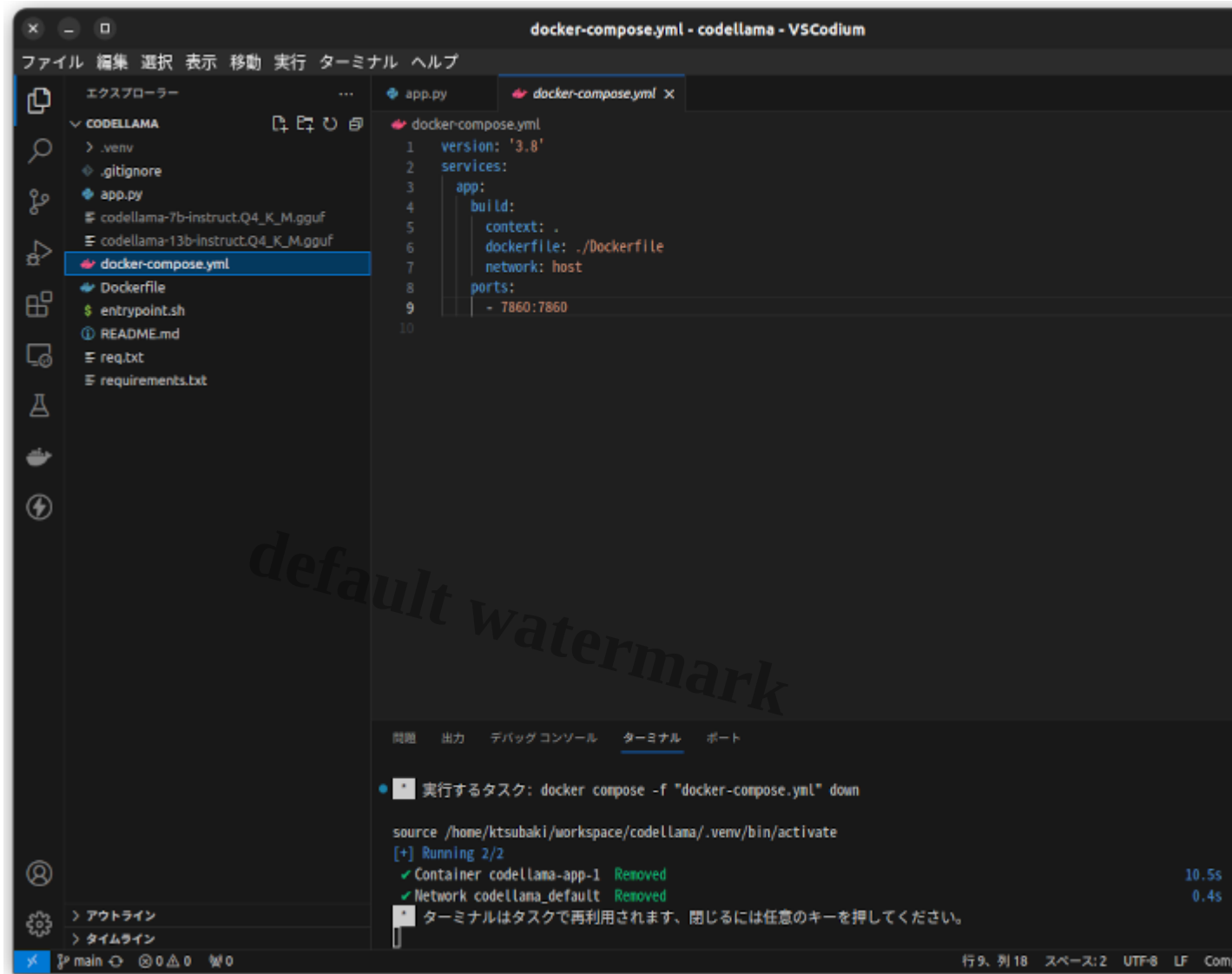
EXPLORER
CODELLAMA
├── .venv
├── .gitignore
├── app.py
├── codellama-7b-instruct.Q4_K_M.gguf
├── codellama-13b-instruct.Q4_K_M.gguf
├── docker-compose.yml
├── Dockerfile
├── entrypoint.sh
├── README.md
├── req.txt
└── requirements.txt

app.py
1 import gradio as gr
2 import time
3 from transformers import AutoModelForCausalLM
4
5 def load_llm():
6     llm = AutoModelForCausalLM.from_pretrained(
7         "codellama-13b-instruct.Q4_K_M.gguf",
8         model_type='llama',
9         max_new_tokens = 1096,
10        repetition_penalty = 1.13,
11        temperature = 0.7,
12        gpu_layers=0,
13        context_length=4096
14    )
15    return llm
16
17 def llm_function(message, chat_history):
18
19     llm = load_llm()
20     response = llm(
21         message
22     )
23     output_texts = response
24     return output_texts
25
26 title = "CodeLlama 13B GGUF Demo"
27
28 examples = [
29     'Write a python code to connect with a SQL database and list down all the tables.',
30     'Write the python code to train a linear regression model using Scikit Learn.',
31     'Explain the concepts of Functional Programming.',
32     'Can you explain the benefits of Python programming Language?'
33 ]
34
35 gr.ChatInterface(
36     fn=llm_function,
37     title=title,
38     examples=examples
39 ).launch(server_name='0.0.0.0')
```

Footer Tagline

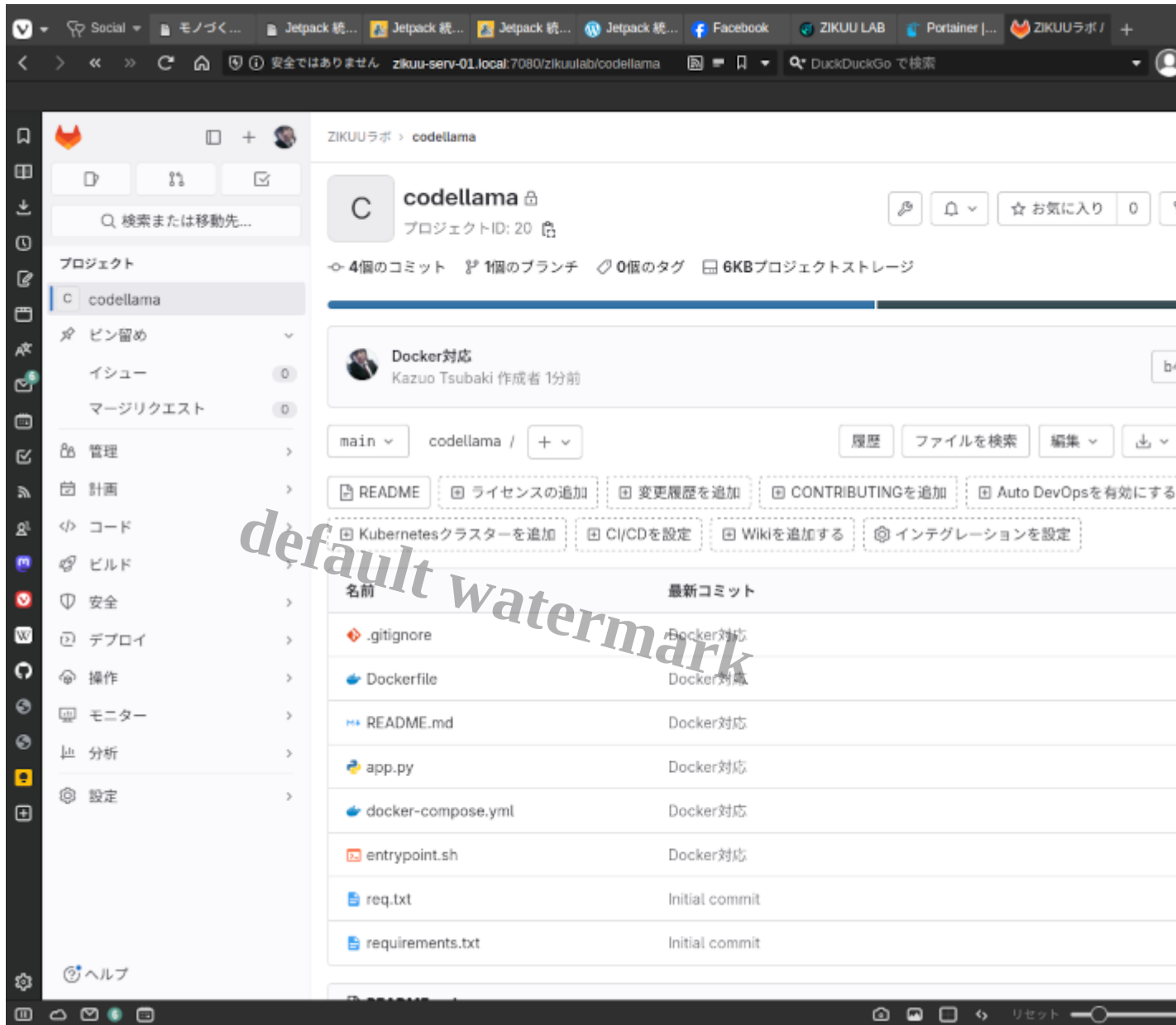


docker compose



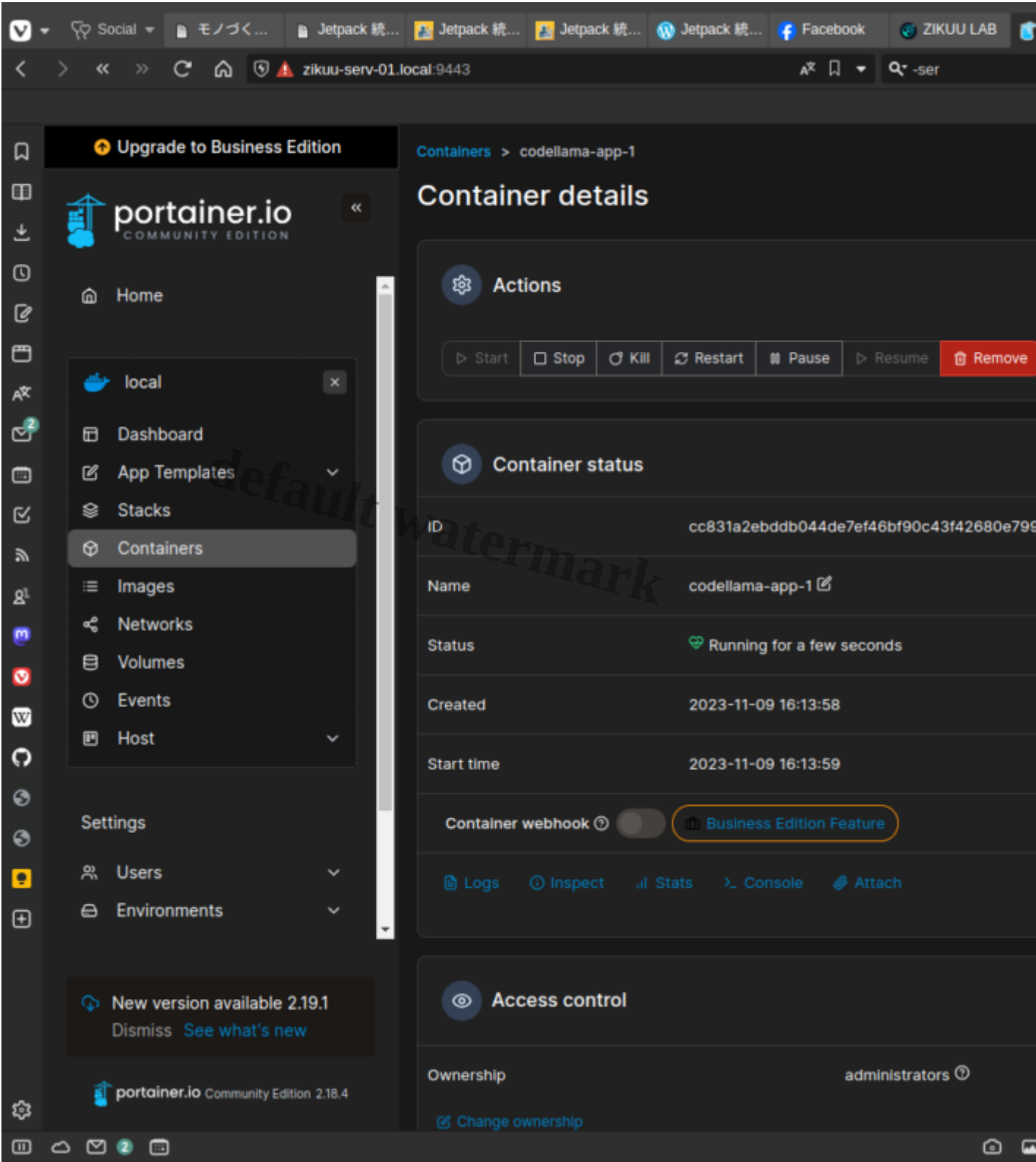
ã?¢ã??ã•¥ã•ã??ã¼ã•sã•ã?µã?¼ã?ã?¼ã•«ã?¢ã??ã?ªã?±ã?¼ã?ã?sã?³ã??ã??ã??ã?ã?ã?ã?ã?ã?
ã?ã?ã?«æ¬¼ã•@æ???é ?ã??è,ã•¿ã•¼ã?ã??ç?¾æ??ç?¹ã•sã•ã?ã??ã?ã?æ?æ°ç??ã?ªã??ã??æ?¹ã•sã•
ã???

i¼?i¼ã?ã?ã?ã?ã?,ã?sã?ã??ã??ã¼ã?µã?¼ã?ã?¼ã?@GitLabã•«ã??ã??ã?ã?ã?ã?ã?

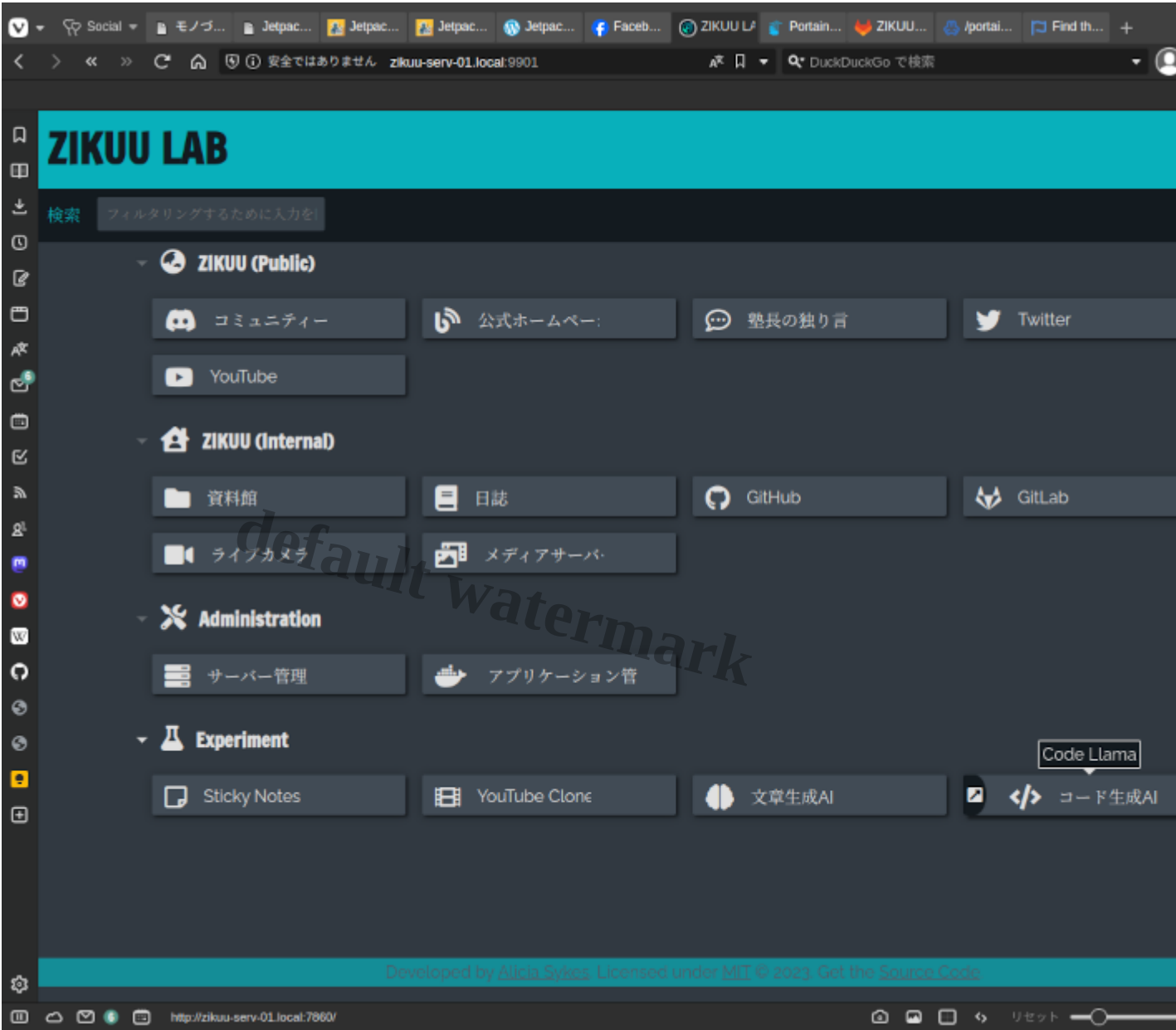


i¼?i¼?ã?µã?¼ã?•ã?¼ã«ã?ã?°ã?²ã³ã•?ã|ã??ã?,ã?§ã?¬ã??ã??ã?¬ã?ã?¼ã³ã?ã|i docker
compose up -d ã?ª®?èj?ã?ª²§ã•ã•ã?£ã?ã?«ã?ã?jã?²ã?«ã?ã?ã?ã?|ã³ã?ã?¼ã??ã•?ã??ã•®ã•
§Dockerã?²ã?jã?¼ã?,ã?ã½?ã??ã??ã??ã•¾ã•§ã«æ???é??ã•ã?ã?ã?ã??ã?¾ã?ã??

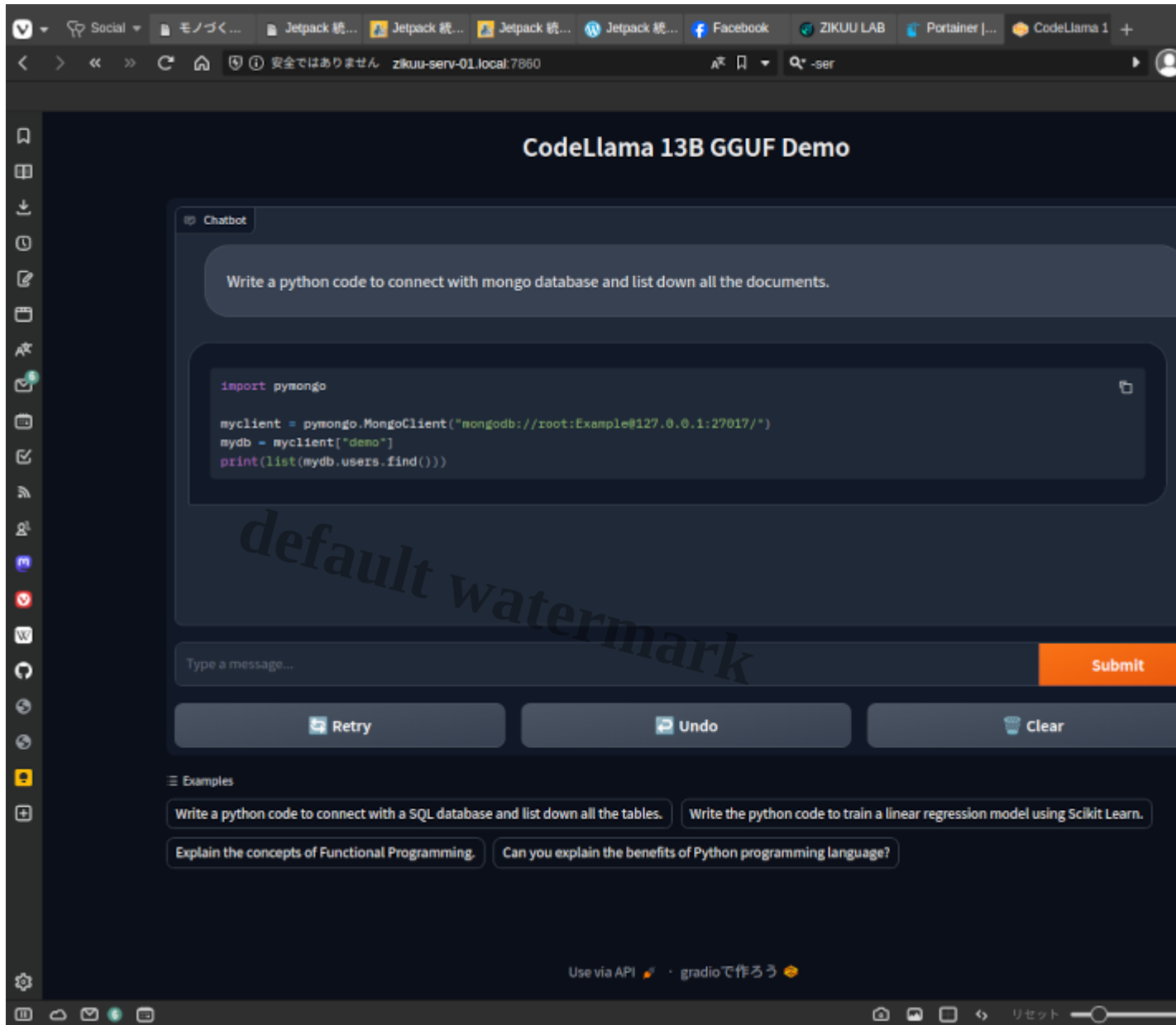
ï¼¼?¼¼?ã³ã³ã??ã??ã??èµ.â?ã?ã?ã?®ã??Portainerã·Sè|?ã·|ã·ã??ã·ã·ã?ã·ã·Sã·ã??



i¼?i¼?Dashyã•@ã??ã??ã?·ã?¥ã??ã?¼ã??ã?«ã?¢ã??ã?ªã?±ã?¼ã?·ã?§ã?³ã??ç?»é?²ã?ã?¾ã?ã??

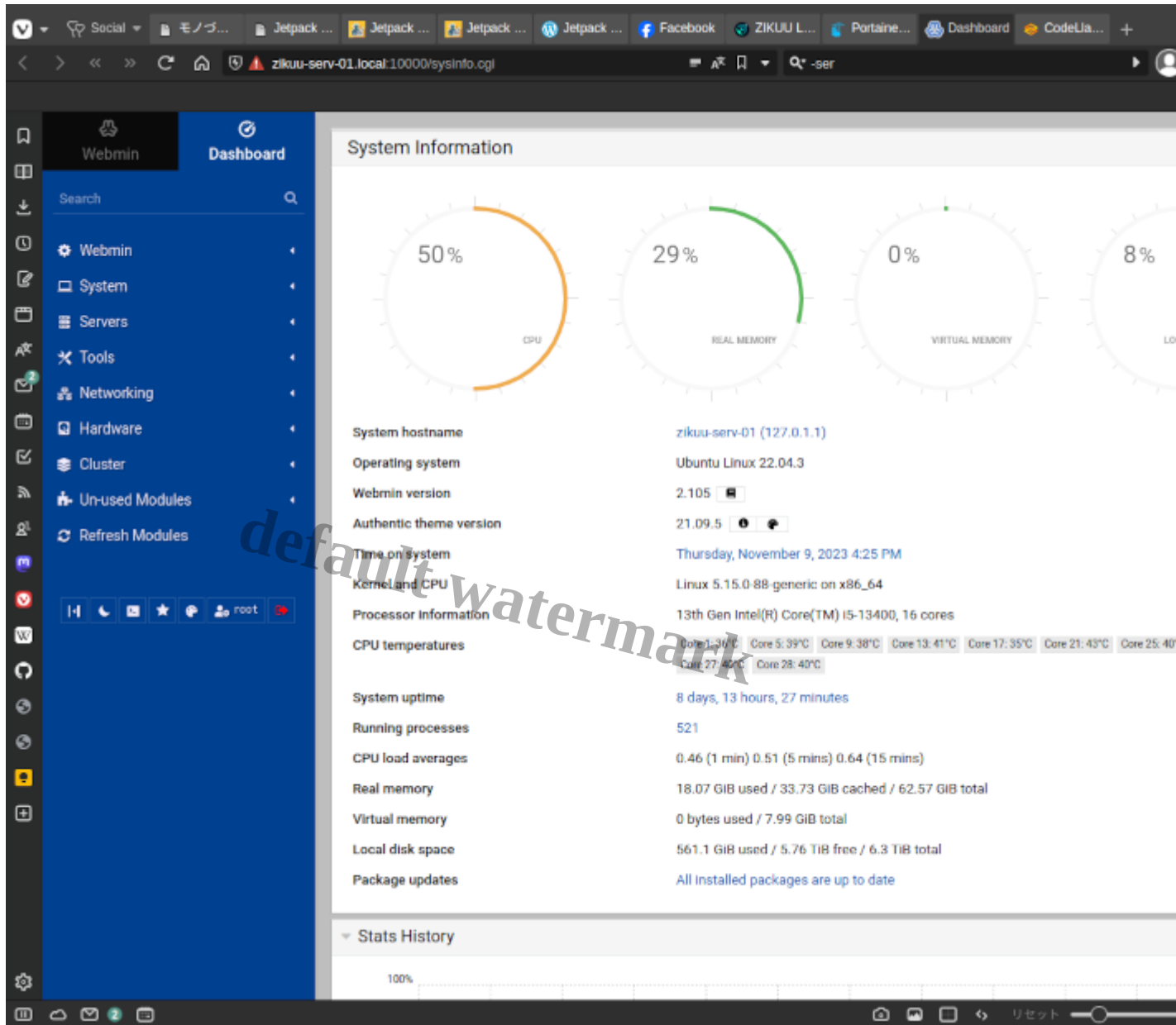


i¼?i¼?ã?ã??ã?ªã?±ã?¼ã?·ã?šã?³ã???é,ã??ã•sèµ·å??ã?ã?ã•Gradioã•šã½?ã•£ã•?Web UIã•é??ã
•ã•¼ã•?ã??



ã•ã?ã?ã?ã?ã?ã?ã^aæµ•ã??ã•šã?ã??ã?³ã?ã?³ã??ã??ã ¥å??ã?ã??ã•®ã•¯¼ã?é?¨ã??ã•ã?ã•šã•
?ã??

ã?³ã?¼ã??ç??æ?•ä,ã•®ã?µã?¼ã?•ã?¼ã•®æ§?åã•§ã•?ã??CPUä½ç?´ç??ã?•52%ã•ã??ã•?ã•¾ã•
§ä,?ã•?ã??ã•¾ã•?ã??



ã•?ã??ã•sã??æ??ã«æ¯?ã¹ã??ã°é???ç?ºã•?ã??é•?ç?`ã¾ã•sã•®â¥ç¨?ã•?é??ã•œ¥½ã«ã•ã•£ã|ã•?ã•
¾ã•?ã•?ã•?œ??çµ?ç?ªã«ã¯GlTLabã«ã??ã??ã•·ã¥ã•?ã•?ã??è?ªâ??ç??ã«ã?µã?¼ã•?ã?¼ã•
«ã??ã??ã?ã?ªã•?ã??ã??ã??ã•?ã«ã•?ãªã•ã|ã¯ã?ã•?ã•¾ã•?ã??ã??

Category

1. $\neg \neg \neg \neg \neg$

Tags

1. AI
2. CTransformers
3. Dashy
4. Docker
5. Gradio

-
- LLM
 - Portainer
 - Python
 - Webmin
 - åðšè!•æ¨jè¨?èª?ã?¢ã??ã?«

Date Created
2023å¹11æ??13æ?¥

Author
kazuo-tsubaki

default watermark