



LLM Broker•®OpenAI API•«ã?¹ã??ã?ªã?¼ã? å?ºå??æ©?è?½ã??å®?è£ã?ã??

Description

æ?·æ?¥ã·-ã?Open WebUIã·ã??OpenAI APIã·\$LLM Brokerã«æ?¥ç||?ã·\$ã·ä??ã?·ã·ã·ã·ã·ã·ã·
ã·¾ã·\$ã??ã??ã·¾ã·ã·ã·?
ä»?æ?¥ã·-ã?LLM Brokerã«ã¹ã??ãªã¹¼ã? å°ºã??æ©?è½ã??å®?è£ ¤ã·¾ã·ã·ã?

é??ç?°ã•\$ã½¿ã•£ã•!ã•?ã?PCã• CPUã•Core i5 13500ã•GPUã•?RTX A4000 16GBã•\$ã•?ã??

ã•ã??ã•\$llama.cppã?µã?¼ã?•ã?¼ + gpt-oss-20b_Q4_K_M.ggufã??èµ°ã??ã?•ã?ã?•LLM Brokerã?
?ã??æ?¥ç¶?ã?•ã??ã?•ã?•ã®?ã®?ç??ã?«90ã??115ã??ã?¼ã?¬ã?¾ç§?ã?ã??ã?•ã?•ã®ç??æ?•é??å¶ã?
?å°°ã?¶ã?•ã?¾ã?•ã??



ã¼ã•ã??ã?¼ã?«ã?¼ã•ã?°ã•?æ©?è?½ã•®ã®?è£ã•?ã•§ã•ã•|ã•?ã•ã•?ã•®ã•§ã®?æ?•ã•§ã•¬ã?ã??ã•¼ã•?ã??ã•?ã•?ã•?ã•?ã•?ã•?é?ã¼µã??ã•°Ollamaã•«è¿ã•?æ©?è?½ã•?ã®?ç¼ã•§ã•ã•ã•?ã•§ã•?ã??

é??ç?°ã•®llama.cppã•®è"ã®?ã•§ã•?ã??

gpt-oss:20bã•§æ?±ã•?ã??ã??ã??ã??ã?"ã?¹ã??é?•ã?ã•£ã•±ã•?ã•«è"ã®?ã•?ã•|ã•?ã•¼ã•?ã??

```
services:
  llama-text:
    image: ghcr.io/ggml-org/llama.cpp@sha256:13f61752307fc4b96c8607a1bc03f977a
    container_name: llama-text
    restart: unless-stopped

ports:
  - "8081:8080"

volumes:
  - ./models:/models:ro

command:
  - -m
  - /models/gpt-oss-20b-Q4_K_M.gguf
  - --host
  - 0.0.0.0
  - --port
```

```
- "8080"
- --n-gpu-layers
- "99"
- --ctx-size
- "131072"
- --cache-type-k
- q8_0
- --cache-type-v
- q8_0
- --flash-attn
- "on"
- --batch-size
- "2048"
- --ubatch-size
- "512"
- --temp
- "1.0"
- --top-k
- "64"
- --top-p
- "0.95"
- --min-p
- "0.0"
- --repeat-penalty
- "1.0"
- --threads
- "4"
- --parallel
- "1"
- --cont-batching
- --log-verbose
- --mlock
```

default watermark

```
ulimits:
  memlock:
    soft: -1
    hard: -1
```

```
deploy:
  resources:
    reservations:
      devices:
        - driver: nvidia
          device_ids:
            - "0"
          capabilities:
            - gpu
```

ã?³ã?³ã??ã?¬ã?¹ã??ã?µã?²ã?°128Kã\$ã?•ã?ã??ã?ã?ã?ã?@æ?Sè?½ã?¿ã?°ã??ã?°ã?é?ã?
?ã??æ?Sè?½ã?\$ã?ã??

å?¥ã?®ã??ã?£ã?é•ã?PCi¼?ã?ã?¶ã??ã?ã?®é??ç?°æ©?ã?®å?¿ã??ã?ã?ã?ã?®é??å?¼?ã?ã?¬ã?
PCIe 3.0æ?¥ç¶?ã?®RTX 3060ã\$gemma4 E4Bã??è¼?ã?ã?¶ã?ã?¾ã?ã?ã?ã?¼¼ã?ã??ã?ã?æ"å®?ã?
§70ã??ã?¼ã?¬ã?³/ç§\$?ã?ã??ã?¿?°ã?¶ã?ã?¾ã?ã??

Category

1. æ?¥ã? ã•®æ´»å??

Tags

1. AI
2. LLM
3. LLM Broker
4. ä,?é??å??å,?
5. å±±±æ€¨ç??

Date Created

2026å¹´7æ??30æ?¥

Author

kazuo-tsubaki

default watermark