LLM Brokerã??ä½?ã??

Description

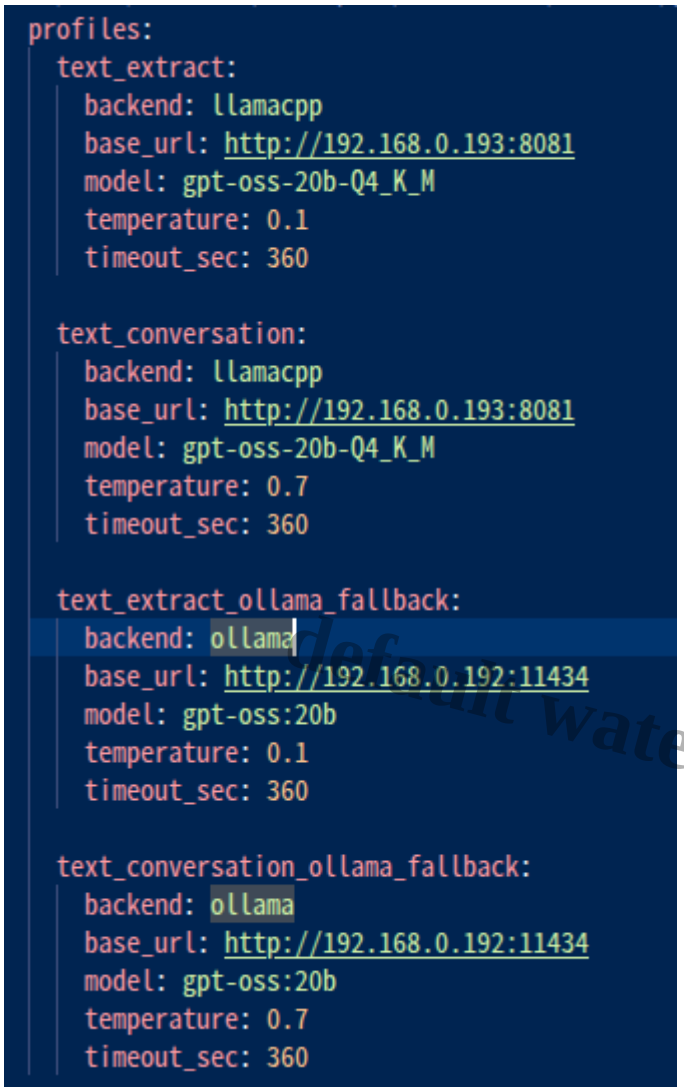
è¤?æ?°ã•®LLM Runner ã??ç?¨é??ã•?ã•¨ã•«å??ã??æ?¿ã•?ã•¦ä½¿ã•?ã??ã??ã•?ã•«ã•?ã•?ã•?ã•¨æ?•ã•?ã?•LLM Brokerã¨ã•?ã•?ã?µã?¼ã?•ã?¼ã•®é??ç?ºã??å§?ã?•ã•¾ã•?ã•?ã??

æ?¬ç?ªç?°å¢?ã•®ã?µã?¼ã?•ã?¼ã•¯ã?•ã•?ã•?ã??ç?•é?»å??ã•ªã??ã??PCã•«ã•?ã??ã•?ã•¨æ?•ã•£ã•¦ã•?ã•¾ã•?ã•?ã?•Strix Pointæ•è¼?ã??ã??PCã•®ã??ã•?ã•ªdGPUã•?è¼?ã•£ã•¦ã•?ã•ªã•?PCã•§LLMã??èµ°ã??ã•?ã??ã•?ã•¨æ?•ã•?ã•¨ã?•ç•¾æ??ç?¹ã•§ã•¯llama.cppä¸?æ??ã§ã•?ã??

è©¦ã•?ã•«ã?•Xeon W-1250P (6C/12T)ã•¨RTX 3060ã??æ•è¼?ã•?ã•?PCã•¨ã?•Core i5 13500 ã•¨ RTX A4000ã??æ•è¼?ã•?ã•?PCã•§llama.cppã?µã?¼ã?•ã?¼ã??å??ã•?ã•?ã•¦ã?•gpt-oss-20b-Q4_K_M.ggufã??ä½¿ã•£ã•¦ã•¿ã•?ã??ã?•ã••ã??ã•?ã??80 tok/secã?•100 tok/secã•¨ã•?ã•?ç??æ?•é??åº¦ã??å¾?ã??ã??ã•¾ã•?ã•?ã??

llama.cppã•ªã??ã?•Strix Pointæ•è¼?ã??ã??PCã•§ã??å®?ç?¨ç??ã•ªé??åº¦ã•?å¾?ã??ã??ã??ã•?ã??ã•ªã•?ã•¨ã•?ã•?ã•?ã•¨ã??ã•?ã•£ã•¦ã•?ã••ã??ã•¾ã•§é??ç?ºã•?ã•¦ã••ã•?ã?¢ã??ã?ªã?±ã?¼ã?·ã?§ã?³ã•¨ä»?å¾?é??ç?ºã•?ã??ã?¢ã??ã?ªã?±ã?¼ã?·ã?§ã?³ã•§ã?•ç?¨é??ã•«å¿?ã•?ã•¦ã?•LLM Runnerã??å??ã??æ?¿ã•?ã??ã??ã??ã??ã•?ã•«ã•?ã•¦ã•?ã••ã•®ã•?è?¯ã?ã•¨ã•?ã•?å?¤æ?§ã?•ã•?ã•®LLM Brokerã•®é??ç?ºã•«è?³ã•£ã•?ã?•ã•?ã•§ã•?ã??

ã•?ã??ã•ªé¢¨ã•«ã??ã?ã??ã?£ã?¼ã?«ã??å®?ç¾©ã•?ã•¾ã•?ã??



```yaml
profiles:
  text_extract:
    backend: llamacpp
    base_url: http://192.168.0.193:8081
    model: gpt-oss-20b-Q4_K_M
    temperature: 0.1
    timeout_sec: 360

  text_conversation:
    backend: llamacpp
    base_url: http://192.168.0.193:8081
    model: gpt-oss-20b-Q4_K_M
    temperature: 0.7
    timeout_sec: 360

  text_extract_ollama_fallback:
    backend: ollama
    base_url: http://192.168.0.192:11434
    model: gpt-oss:20b
    temperature: 0.1
    timeout_sec: 360

  text_conversation_ollama_fallback:
    backend: ollama
    base_url: http://192.168.0.192:11434
    model: gpt-oss:20b
    temperature: 0.7
    timeout_sec: 360
```


å?¼ã•³å?ºã•?å•´ã•¯ã?LLM Brokerã•«å¯¾ã?ã•¦ã?•ã??ã?ã??ã?£ã?¼ã?«å••ã•§LLMã??å?¼ã•³å?ºã•?ã??ã??ã•?ã•«ã•ªã??ã•¾ã•?ã??
è¦•ã•?ã??ã•«ã?•ã?¢ã??ã?ªã?±ã?¼ã?·ã?§ã?³ã•¨LLM Runnerã??ä¸ç¶?ã•?ã??è??ã•?å±¤ã??ä½?ã•£ã•?ã?•ã•?ã•§ã•?ã??

ä¸?ã•®ã?¹ã?¯ã?ªã?¼ã?³ã?·ã?§ã??ã??ã¯ã?ã??ã?¹ã??ç?¨ã•®ã?¹ã?¯ã?ªã??ã??ã§ã?•ï¼?ã•¤ã•®ã??ã?-ã??ã?£ã?¼ã?«ã?•ï¼?ã•¤ã•¯llama.cppã??ä½¿ã?ã??ã•®ã?•ã??ã•?ï¼?ã•¤ã•¯Ollamaã??ä½¿ã?ã??ã•®ã?•ã•§LLMã??å?¼ã•³å?ºã•?ã•?ã•¨ã••ã•®çµ•æ??ã•§ã•?ã??

ã•¡ãªã•¿ã•«ï¼?ã•¤ã?•ã•¯Xeon W-1250P + RTX 3060ã?ï¼?ã•¤ã?•ã•¯Core i5 13400 + RTX 2000 Adaã??æ•è¼?ã•?ã•?PCã•§å??ã•?ã•¦ã•?ã??LLM Runnerã•?ã??ã•®å¿?ç?ã•§ã•?ã??

Category

1. æ?¥ã? ã•®æ´»å??

Tags

1. AI
2. llama.cpp
3. LLM

4. LLM Broker
5. Ollama
6. ä¸?é??å??å¸?
7. å±±æ¢¨ç??

Date Created
2026å¹´7æ??4æ?¥
Author
kazuo-tsubaki