



LLM Runtime Agent

Description

LLM Runtime Agent は、ZIKUU Mini 上で動作する LLM / AI 増強バックエンドの起動・停止・状態確認を統一するための小さなホスト常駐サービスです。

LLM Broker 自身は Docker コンテナとして動作するため、ホスト上の `systemd` や `Docker Compose` を直接操作せず、ホスト側の Runtime Agent にライフサイクル管理を委譲します。

現在は以下のバックエンドに対応しています。

- llama.cpp
- Lemonade Server

LLM Broker は、ZIKUU Mini 上で動作する LLM / AI 増強バックエンドの起動・停止・状態確認を統一するための小さなホスト常駐サービスです。

LLM Broker 自身は Docker コンテナとして動作するため、ホスト上の `systemd` や `Docker Compose` を直接操作せず、ホスト側の Runtime Agent にライフサイクル管理を委譲します。

現在は以下のバックエンドに対応しています。

- llama.cpp
- Lemonade Server

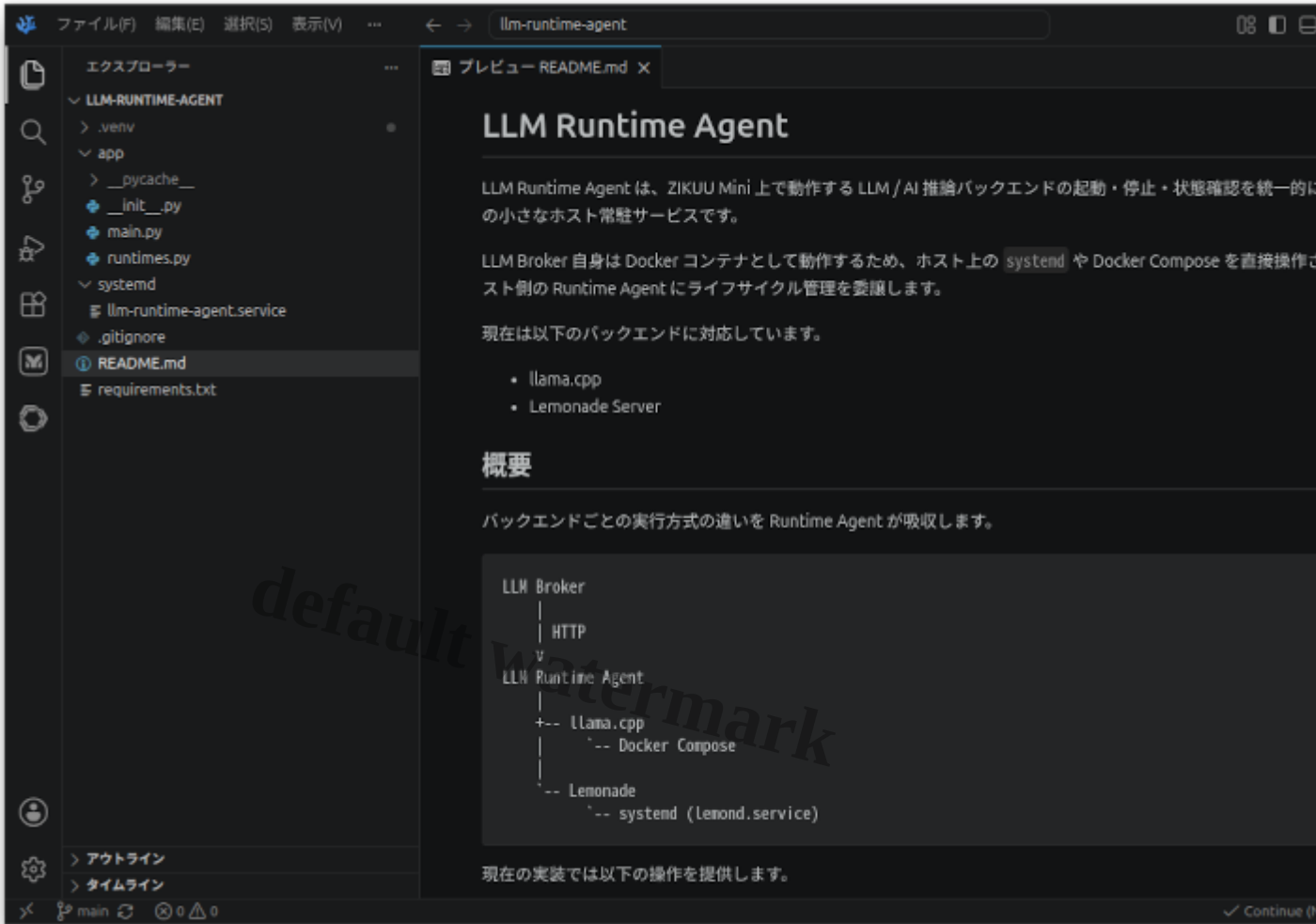
LLM Runtime Agent は、ZIKUU Mini 上で動作する LLM / AI 増強バックエンドの起動・停止・状態確認を統一するための小さなホスト常駐サービスです。

LLM Broker 自身は Docker コンテナとして動作するため、ホスト上の `systemd` や `Docker Compose` を直接操作せず、ホスト側の Runtime Agent にライフサイクル管理を委譲します。

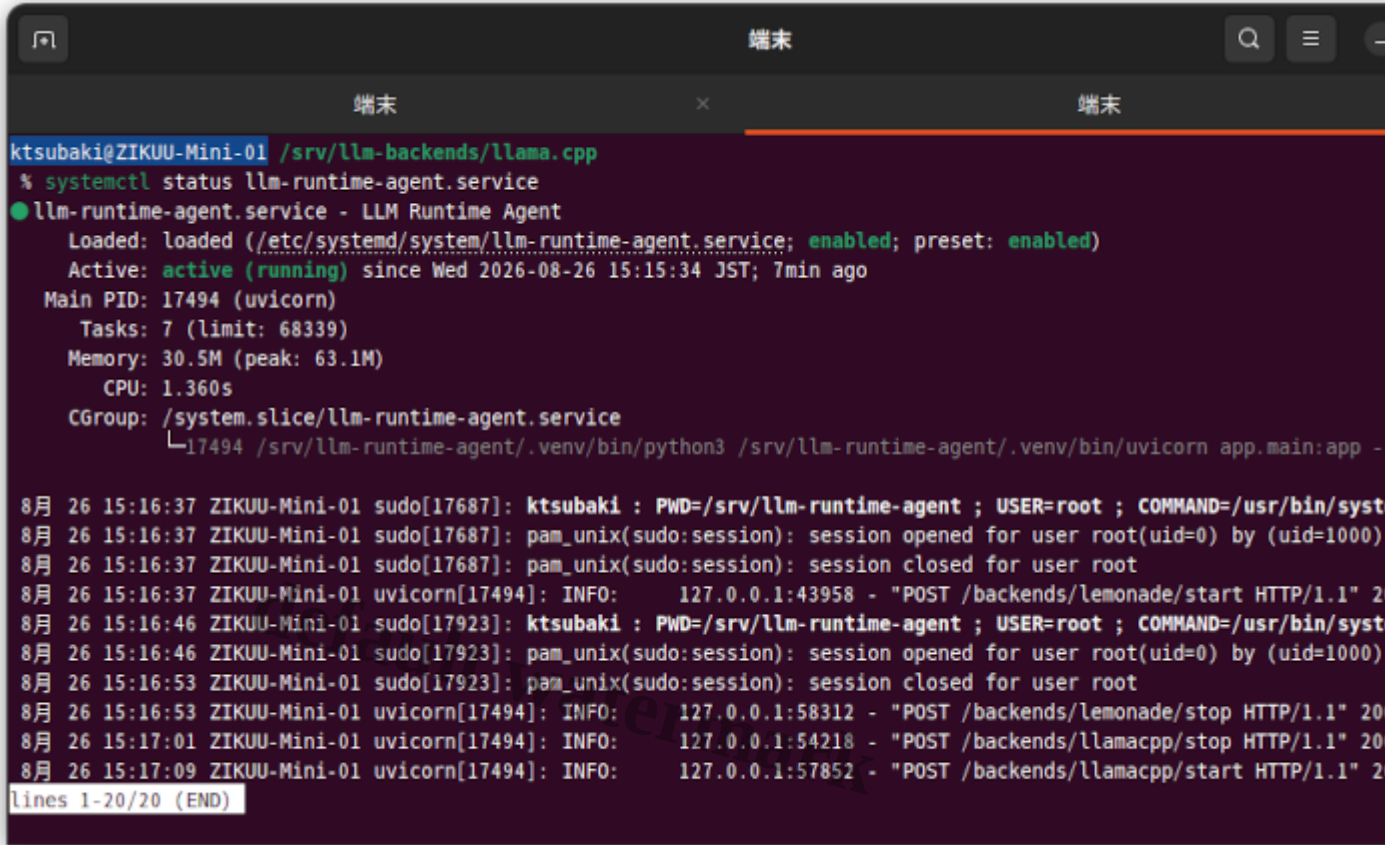
現在は以下のバックエンドに対応しています。

- llama.cpp
- Lemonade Server

- llama.cpp
- Lemonade Server



LLM Runtime Agentは、systemd上で動作する LLM / AI 推論バックエンドの起動・停止・状態確認を統一的に提供するための小さなホスト常驻サービスです。



ã•?ã??ã•?ã?•Dockerã?³ã³ã??ã??ã•ã•?ã•?ã?ã?ã•llama.cppã?µã?¼ã?•ã?¼ã•systemdã•
®ã?µã?¼ã??ã?¹ã•ã•?ã•?ã?ã?ã•Lemonadeã•®èµ•ã??ã?»ã•?æ£ã??èj?ã•?ã•¾ã•?ã??

```
终端
ktsubaki@ZIKUU-Mini-01 /srv/llm-runtime-agent
% curl -s http://127.0.0.1:8090/backends/llamacpp/status | jq
{
  "backend": "llamacpp",
  "status": "running"
}
ktsubaki@ZIKUU-Mini-01 /srv/llm-runtime-agent
% curl -s http://127.0.0.1:8090/backends/lemonade/status | jq
{
  "backend": "lemonade",
  "status": "stopped"
}
ktsubaki@ZIKUU-Mini-01 /srv/llm-runtime-agent
% curl -s -X POST \
  http://127.0.0.1:8090/backends/lemonade/start | jq
{
  "backend": "lemonade",
  "status": "started"
}
ktsubaki@ZIKUU-Mini-01 /srv/llm-runtime-agent
%
```

ç?¶æ ?ã??è|?ã??ã?ã?? status ã??å?¼ã?ã?¾ã?ã??

```
# llama.cppã?µã?¼ã?ã?¼ã?@èµ·å??ç?¶æ??ã??è|?ã??
curl -s http://127.0.0.1:8090/backends/llamacpp/status
```

```
# lemonadeã?µã?¼ã?ã?¼ã?@èµ·å??ç?¶æ??ã??è|?ã??
curl -s http://127.0.0.1:8090/backends/lemonade/status
```

èµ·å??ã?ã?æ¢ã?ã?start/stopã??å?¼ã?ã?¾ã?ã??

```
# lemonadeã?µã?¼ã?ã?¼ã?@èµ·å??
curl -s -X POST \
  http://127.0.0.1:8090/backends/lemonade/start | jq
```

```
# lemonadeã?µã?¼ã?ã?¼ã?@å?æ¢
curl -s -X POST \
  http://127.0.0.1:8090/backends/lemonade/stop | jq
```

```
# llama.cppã?µã?¼ã?ã?¼ã?@èµ·å??
curl -s -X POST \
  http://127.0.0.1:8090/backends/llamacpp/start | jq
```

```
# llama.cppã?µã?¼ã?ã?¼ã?@å?æ¢
curl -s -X POST \
  http://127.0.0.1:8090/backends/llamacpp/stop | jq
```

ã?ã??ã-ã?æ???çµ?ç??ã«ã?•LLM Brokerã?•æ?“è«?æ©?è?½ã•®è² è•å??æ?£ã?•ã?-
ã?¥ã?¼ã?□ã?³ã?°ã?•ã?¹ã?±ã?,ã?¥ã?¼ã?ªã?³ã?°ã?ã?•©ã??ã?§ã?•ã??ã?°ã?ã?ã?•ã?•ã?•ã?•®ä,?æ-
©ã?§ã?•ZIKUU miniå°?ç?°ã?®ã?µã?¼ã?°ã?¹ã?§ã?•ã?ã?°ã?¾ã?•ã?°ã??

ã?ã•®æ??ç`¿ã•-ã?Ryzen AI NPUã•®æ?`è«?è?½å??ã?æ?è¼¼ã•ã??ã?ã•®ç¶¶ã•ã•šã•ã??

Category

1. æ?¥ã? ã•®æ'»å??

Tags

1. AI
2. Lemonade
3. llama.cpp
4. LLM
5. LLM Broker
6. LLM Runtime Agent
7. ZIKUU Mini
8. ä,?é??å??ä,?
9. å±±æ¢“ç??

Date Created

2026å¹¸æ??26æ?¥

Author

kazuotsubaki