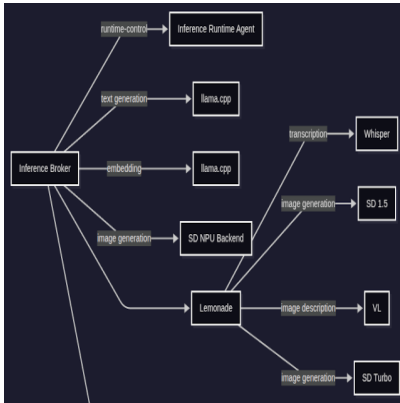




LLM Broker/Vision Broker Inference Broker

Description

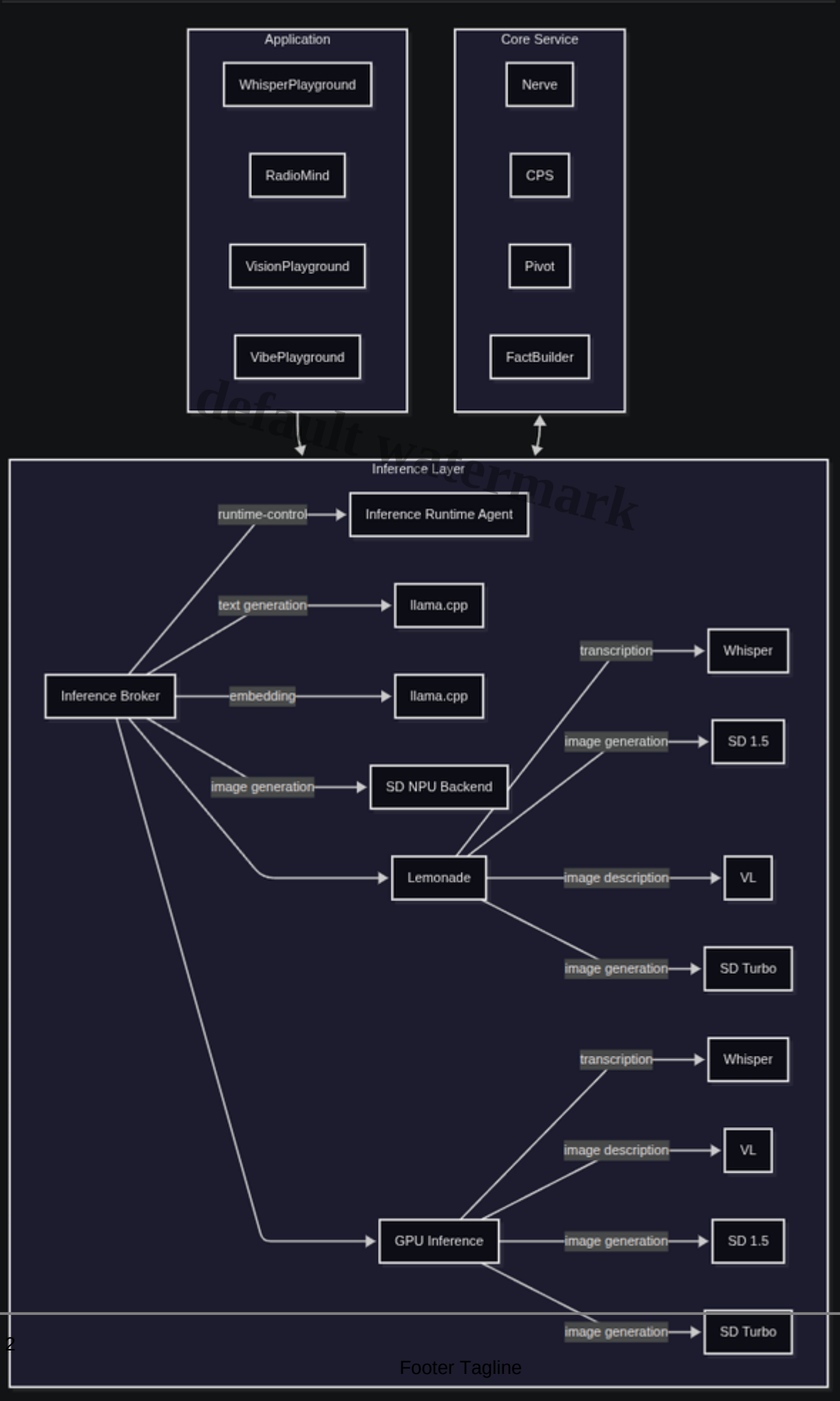
The LLM Broker is a central component that manages the execution of various AI models. It acts as a broker, routing requests to the appropriate backend services. The Vision Broker is a specialized component that handles image-related tasks, such as image generation and image description. The Inference Broker is the main component that manages the execution of various AI models, including LLMs and vision models.

Application The LLM Broker is used to manage the execution of various AI models. It acts as a broker, routing requests to the appropriate backend services. The Vision Broker is a specialized component that handles image-related tasks, such as image generation and image description. The Inference Broker is the main component that manages the execution of various AI models, including LLMs and vision models.

The LLM Broker is used to manage the execution of various AI models. It acts as a broker, routing requests to the appropriate backend services. The Vision Broker is a specialized component that handles image-related tasks, such as image generation and image description. The Inference Broker is the main component that manages the execution of various AI models, including LLMs and vision models.

- Inference Broker - Inference Runtime Agent
- VRAM
- OpenAI API
- llama.cpp
- Lemonade
- SD NPU Backend (ZIKUU)
- GPU Inference (ZIKUU)

Inference Broker周り



ZIKUU Miniã• NPUæè¼?æ©?ãªã®ã·Sã?·ç?»å?·ç??æ?·ã?·ç?»å?·è§£æ?·ã?·æ??å?èµ.ã?·ã?·ã?«ã·¤ã
·?ã?·ã?·NPUã??ã½¿ã?·ã?·ã?·ã?·ã?·ã?³ã??ã?·ã?ã?¢ã?¼ã??ã??ã?¹ã??ã?®GPUã??ã½¿ã?·ã?
ã??ã?·ã?·ã?³ã??ã?«æ?¥ç¶¶?ã?·ã??ã?·ã?·ã?¾ã?·ã??

ZIKUUã•®ã??ã?«ã?·ã?ã??ã? ã•®å´å•ã¬ã?NPUæ•è¼¼?æ©?ã??ä½¿ã•£ã|ã•?ã•ã•?ã•®ã•§ã?•
Lemonadeã¬SD NPU Backendã•®ã??ã?¹ã•?ã•?ã??ã•¾ã•?ã??ã??

[illegible]

Category

1. æ?¥ã? ã•®æ'»å??

Tags

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

Date Created

2026å¹9æ??29æ?¥

Author

kazuotsubaki