KubeCon

CloudNativeCon

China 2024

Running LLMs in the Cloud

Miley Fu, WasmEdge

GitHub / Twitter: @mileyfu

https://github.com/WasmEdge/WasmEdge

https://github.com/LlamaEdge/LlamaEdge

My Calling in Open Source Technology

WebAssembly and Rust Meetup (Wasm Empowering AI)

★★★★★ (14)

Beijing, China

390 members · Public group

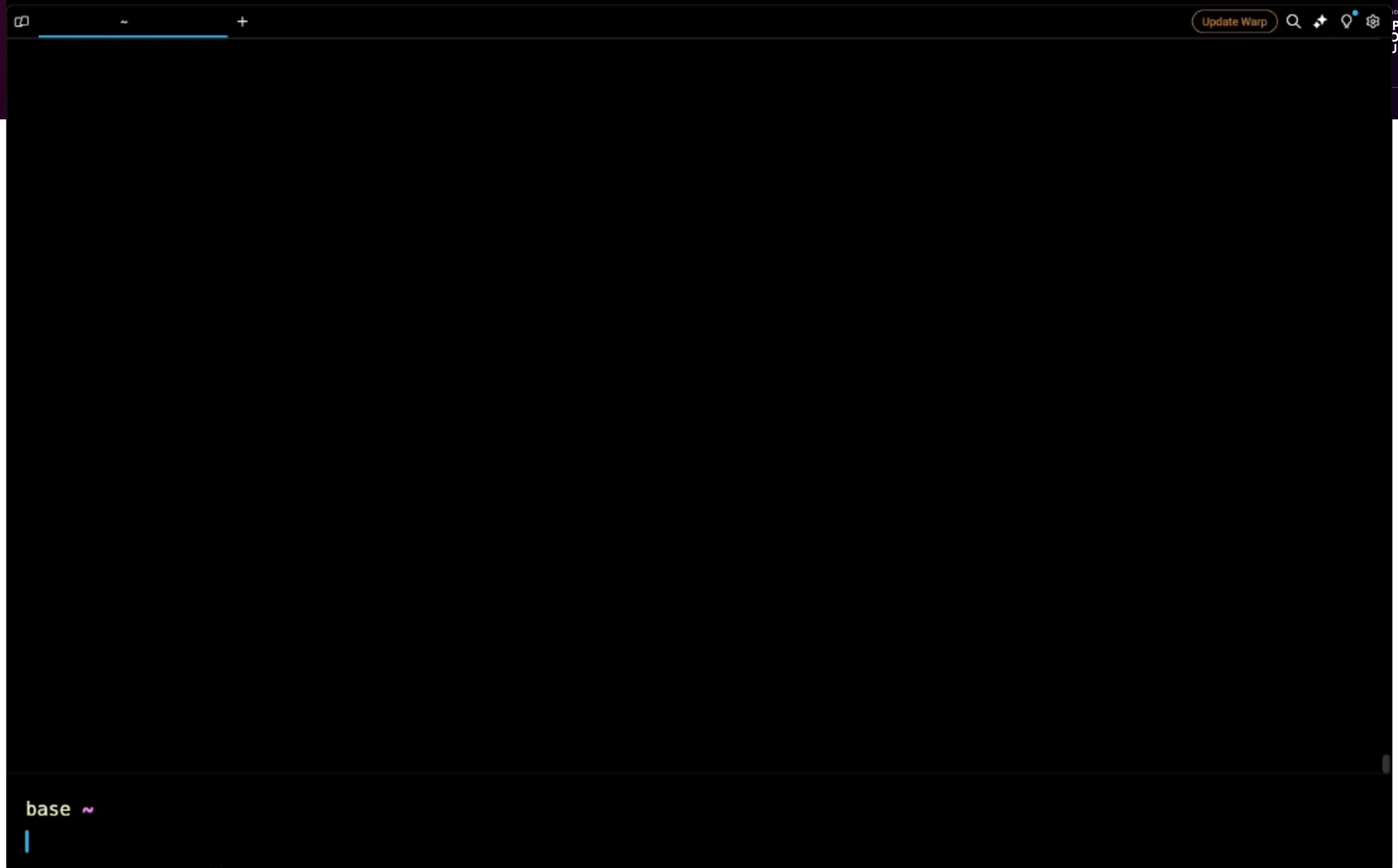
Organized by **Miley Fu** and **2 others**

Embed LLM into your container app


```
docker run --rm -p 8080:8080 --name api-
server secondstate/llama-3-8b-nomic-
1.5:latest ctx size-4096
```

https://hub.docker.com/repository/docker/secondstate/llama-3-8b-nomic-1.5/general

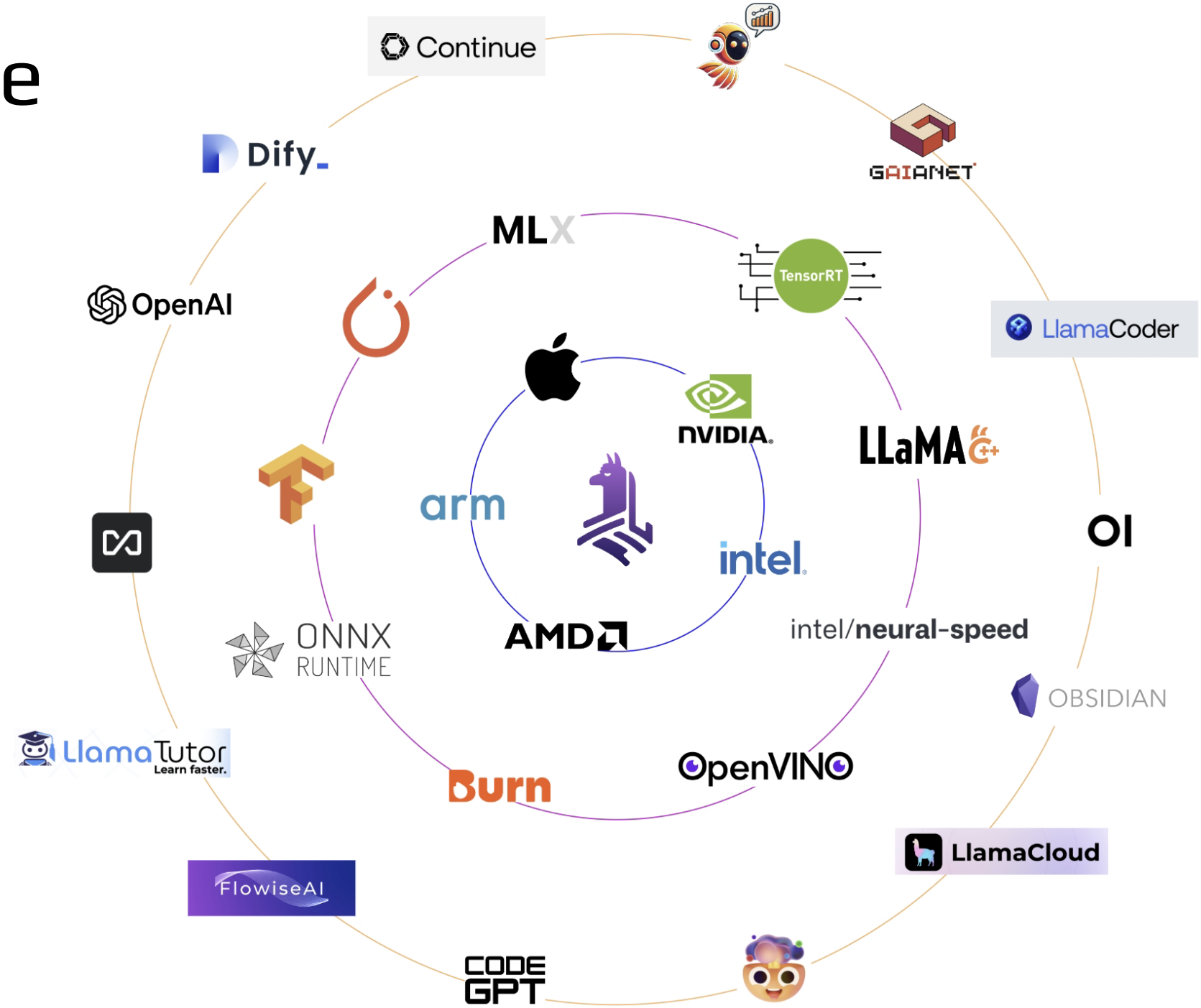
https://github.com/LlamaEdge/LlamaEdge/tree/main/docker


~
Update Warp
base ~

Key features

- Tightly coupled LLM and application
  - Matches prompts, quantization & runtime with the exact version of LLM
  - The container app always works regardless LLM upgrade cycles
- Lightweight
  - Only 5GB as opposed to 10GB PyTorch app
- Portable
  - The same binary app inside the container works on multiple CPUs and GPUs
  - Develop on Mac and deploy on Nvidia
- Easy to embed into Rust / JS / Python apps
- Works with existing container tools, such as K8s

LlamaEdge


Continue
Dify.
GAIANET
MLX
TensorRT
OpenAI
LlamaCoder
NVIDIA
LLaMA C++
arm
intel
OI
ONNX RUNTIME
AMD
intel/neural-speed
OBSIDIAN
LlamaTutor
Learn faster.
Burn
OpenVINO
LlamaCloud
FlowiseAI
CODE GPT

Real-world use cases

1. **Personal LLMs:** Gaia Network, users run personal LLMs with embedded knowledge base.
2. **AI OS:** Open interpreter as a local LLM provider; 51.5k stars
3. **Finance:** Financial analytics bot.
4. **Hardware:** Robot voice control
5. **Education:** UC Berkley TA
6. **Game:** Open-source game engine Cocos AI , use Wasm to run AI models that enhance gameplay experiences enabled by NPC.

Nodes:	28,469
Throughputs:	2884.2M

LlamaEdge: easy LLM deployment + inference

- Single cross-platform binary
  (Automagically take advantage of local hardware accelerators)
- Compile and test apps on one machine (e.g., Mac) and deploy it to another cloud server (e.g., Nvidia CUDA 12)
- The app can be moved around and deploy to new hardware by K8s
- Package the Wasm app into Docker image as an embedded AI / LLM service. Only 1/100 the size of a Python runtime.

LlamaEdge Github

WasmEdge
https://github.com/WasmEdge/WasmEdge

LlamaEdge
https://github.com/LlamaEdge/LlamaEdge

Gaia: The living knowledge server
https://github.com/GaiaNet-AI

Cloud-Native AI Ecosystem

KubeCon | CloudNativeCon | THE LINUX FOUNDATION OPEN SOURCE SUMMIT | AI_dev

China 2024

ters | GROUP: Projects and products | Members | Certified partners and providers | Serverless | Wasm | CNAI | VIEW MODE: Grid | Card | ZOOM: - +

Data Architecture

ClickHouse, druid, cassandra, SCYLLA, HBASE, presto, trino, Spark, Flink, kafka, PULSAR, redis, ALLUXIO, FEAST

Model/LLM Observability

trulens, deepchecks, OpenLIT, DeepFlow

Governance, Policy & Security

JSON, stacklok

General Orchestration

kubernetes CNCF GRADUATED, VOLCANO, ARMADA, RAY

ML Serving

KServe, vLLM

AutoML

HYPEROPT, Katib

Distributed Training

PyTorch, PyTorch, deepspeed, Megatron-LM, Alpa

CI/CD - Delivery

mlflow, MLRun

Vector Databases

Milvus

Data Science

jupyter, PyTorch, dmlc XGBoost, Apache Zeppelin

Workload Observability

Grafana, W&B

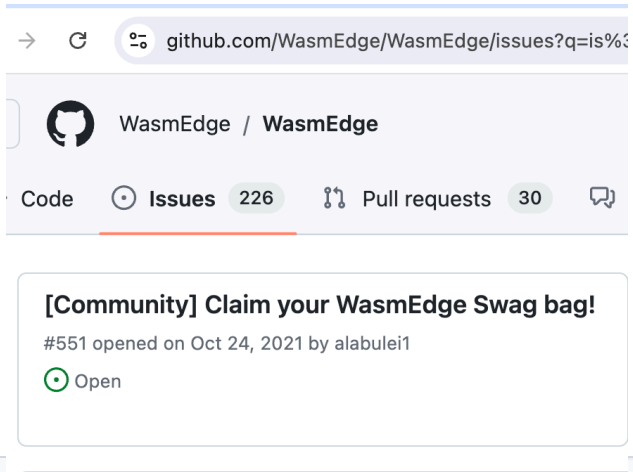
https://landscape.cncf.io/?group=cnai

CNCF + LF AI & Data Survey

How your company uses GenAI tools

- Different use cases
- Modalities/models in use
- Challenges to adoption
- The role of open source in adoption decisions

Calling for contributors



LFX Mentorship

Google Summer of Code

Filters | is:issue is:open label:"LFX Mentorship" | Lab

Clear current search query, filters, and sorts

26 Open | 36 Closed | Author | Label | Projec

- **LFX mentorship (Sept-Nov 2024): WASM Serializer with new proposals** enhancement LFX Mentorship
  #3585 opened 3 weeks ago by q82419
- **LFX mentorship (Sept-Nov 2024): Fix bugs found by fuzzer** enhancement LFX Mentorship
  #3584 opened 3 weeks ago by hydai
- **LFX Mentorship (Sept-Nov 2024): Create an LLM app with deep understanding of a GitHub repo** LFX Mentorship
  #3581 opened last month by juntao
- **LFX Mentorship (Sept-Nov 2024): Create a Wasm-based LLM app for financial analysts** LFX Mentorship
  #3580 opened last month by juntao

YouTube

Second State WeChat

WasmEdge Github

Stay in Touch!

Twitter/ Github

@mileyfu