



WasmEdgeRuntime

# Self-hosted LLMs across all your devices and GPUs

Michael Yuan

x: juntao github: juntao

WasmEdge Runtime: <https://github.com/WasmEdge/WasmEdge>

# All open source

**WasmEdge: The lightweight and cross platform AI runtime**

<https://github.com/WasmEdge/WasmEdge>

**LlamaEdge: The developer platform for LLM apps**

<https://github.com/LlamaEdge/LlamaEdge>

**GaiaNet: The RAG API server and node**

<https://github.com/GaiaNet-AI>

**Demo: The easiest way to  
chat with an open-source  
LLM on your own device**




The background is a solid orange color. In the top right corner, there are several decorative elements: a small orange circle, a larger orange circle with a smaller orange circle inside it, and another small orange circle below the larger one. All these circles have a slight gradient and a small white triangle pointing towards the center.



# Why not just OpenAI?

- One-size fits all
  - Use the largest model for the smallest task
  - Difficulty to finetune models
- Expensive
- Lack of privacy and control
- Censorship and bias



**Marc Andreessen**     
@pmarca



I know it's hard to believe, but Big Tech AI generates the output it does because it is precisely executing the specific ideological, radical, biased agenda of its creators. The apparently bizarre output is 100% intended. It is working as designed.

8:48 AM · 2/26/24 From Earth · **11M** Views

**4.6K** Reposts **481** Quotes **22K** Likes **1K** Bookmarks



# Why LlamaEdge API server?

- Supports a wide range of models on Hugging face (6000+ LLMs)
  - Llama2, Mistral, 01.ai, zephyr, and many many more
  - many embedding models
  - x.ai's Grok
  - advanced MoE and multimodal LLMs
- Supports a wide range devices and drivers. Runs at native GPU speed
  - Nvidia CUDA, TensorRT
  - Apple M chips with metal or MLX
  - Advanced CPUs
  - ARM NPUs
- Multiple models in a single server (embedding and multimodal)
- Forced formatted responses (JSON and function calling)
- Easy to install and run
- Very lightweight – entire runtime + app is less than 30MB



# LlamaEdge is a developer platform

- Build a single **portable** and deployable app
  - Move code closer to model and data
  - Improve efficiency
  - Simplify development and workflow
  - Improve security
- No Python dependency
- Use Rust or JS to extend LlamaEdge components!
- Dev experience that matches the best of OpenAI
  - i.e., highly integrated OpenAI Assistant API



**Greg Brockman** ✓  
@gdb



Much of modern ML engineering is making Python not be your bottleneck.

6:55 AM · 7/6/23 from Earth · **244K** Views



**Santiago Viquez** ✓  
@santiviquez



The best minds of my generation are thinking about how to install Python.



**Chris Albon** @chrisalbon · 1d

What is "the right way" to install Python on a new M2 MacBook? I assume it isn't the system Python3 right? Maybe Homebrew?

3:42 AM · 7/6/23 from Earth · **744K** Views

# Demo: The portable AI inference app in Wasm





## Dev

- Use several different languages to create your apps
  - Currently supports Rust, but JavaScript is almost there.
- Only need to call WasmEdge API to perform inference operations.
  - No need to worry about the GPU drivers or tensor libraries.
- The WasmEdge inference API is based on W3C's WASI NN standard.
- Compile the application to Wasm.
- Distribute and deploy the Wasm binary file using existing tools.





## Ops

- Install WasmEdge with the LLM plugin.
  - It will install GPU drivers and SOTA inference libraries for this device.
- Run the Wasm binary app.
- Bonus: the WasmEdge runtime itself is a security sandbox and can be managed by container tools like K8s, Docker and OpenShift.

**WasmEdge apps for cloud-native!**

**Demo: an integrated  
“assistant API server” built  
on LlamaEdge**

The background is a solid orange color. In the top right corner, there are several decorative elements: a small orange circle, a larger orange circle, and a medium orange circle, all with a slight gradient and a small white triangle pointing towards the bottom right.

# Thank you

Learn more:

<https://github.com/WasmEdge/WasmEdge>

