



Personalized LLM agents on your own devices

Michael Yuan

x: juntao github: juntao

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



Follow along step-by-step

<https://github.com/GaiaNet-AI/workshops/tree/main/20240515-aicamp>

**Demo: The easiest way to chat
with an open-source LLM + a
domain knowledge base on
your own device**



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

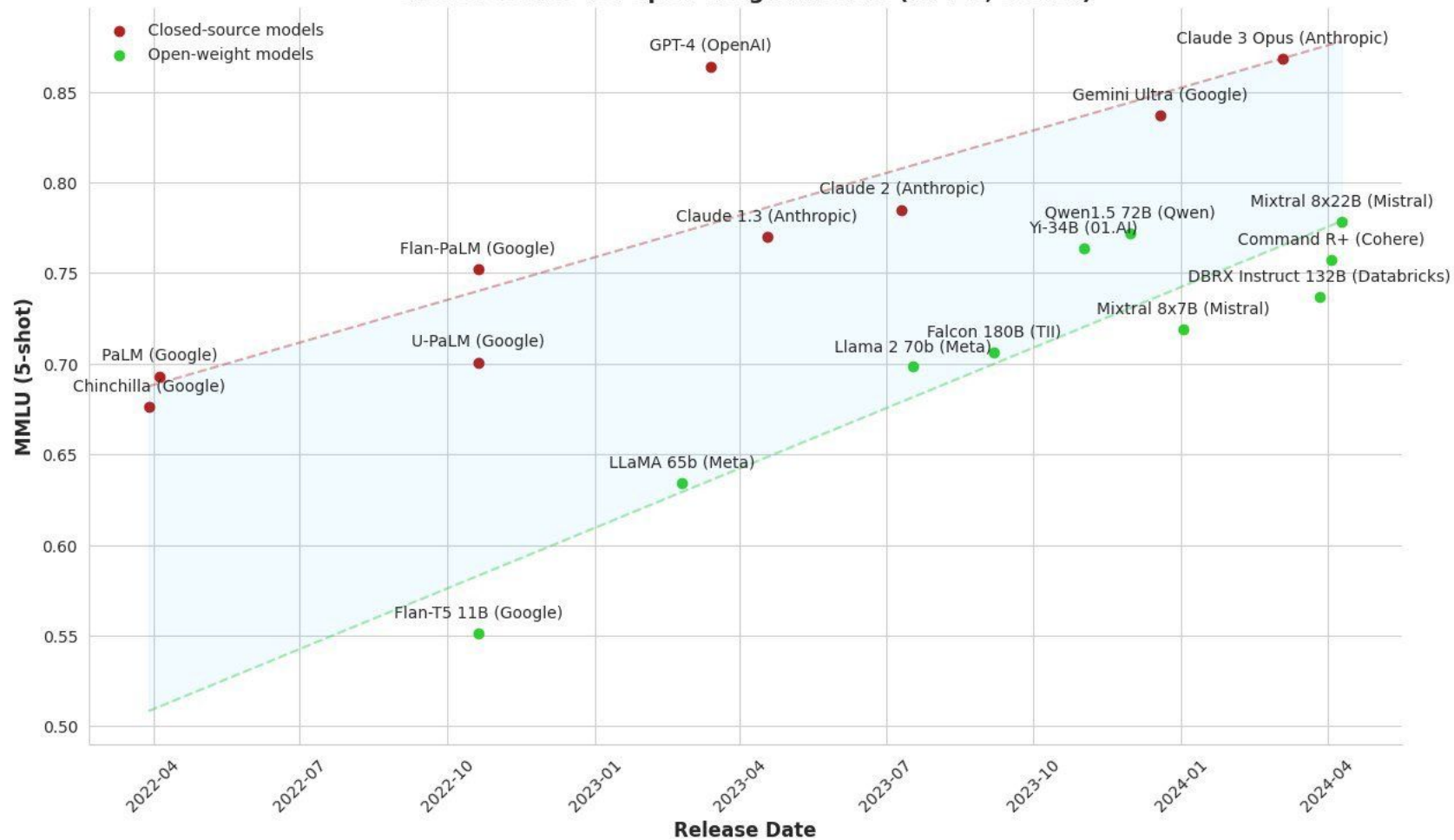
...

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

Closed-source vs. Open-weight models (MMLU, 5-shot)

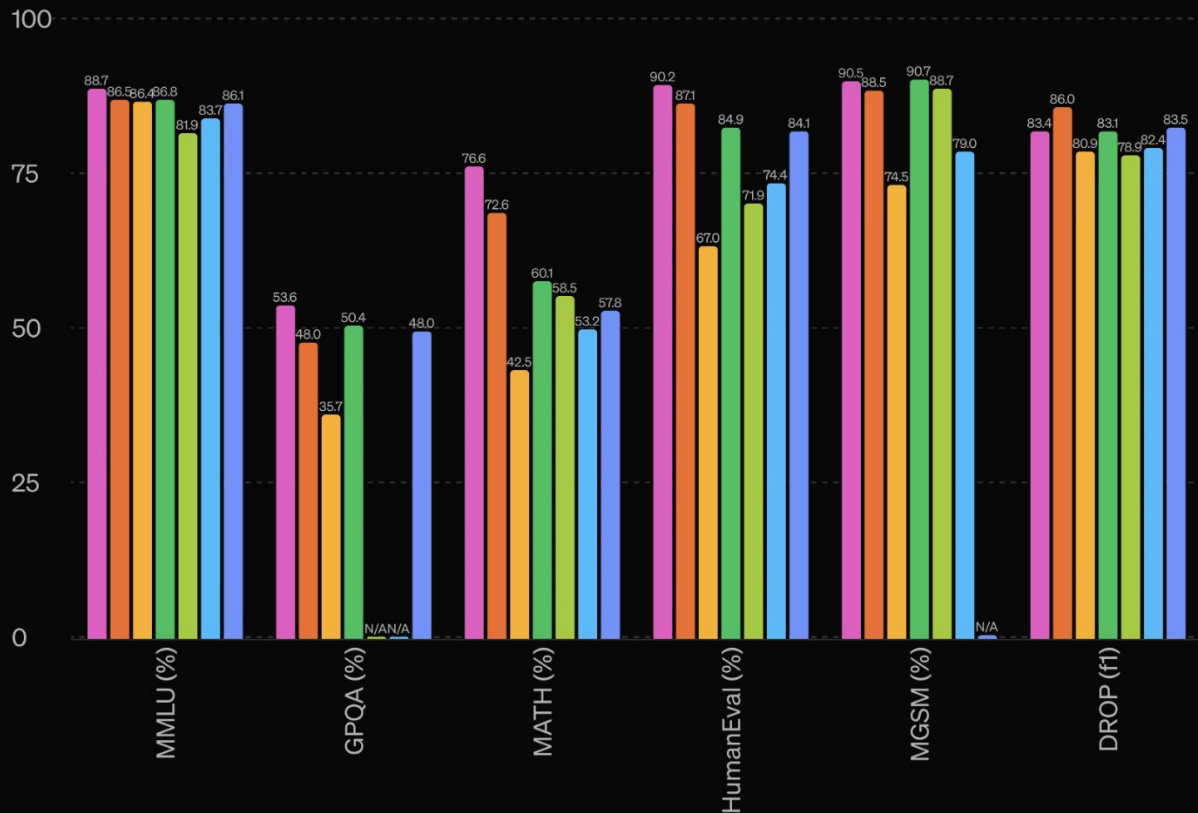


	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	82.0	81.9	79.0
GPQA 0-shot	39.5	41.5 CoT	38.5 CoT
HumanEval 0-shot	81.7	71.9	73.0
GSM-8K 8-shot, CoT	93.0	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	58.5 Minerva prompt	40.5

<https://openai.com/index/hello-gpt-4o/>

Text Evaluation

■ GPT-4o
 ■ GPT-4T
 ■ GPT-4 (initial release 23-03-14)
 ■ Claude 3 Opus
 ■ Gemini Pro 1.5
 ■ Gemini Ultra 1.0
 ■ Llama3 400b



4. Multiagent collaboration



Multiagent Debate

Task	Single agent	Multi-agent
Biographies	66.0%	73.8%
MMLU	63.9%	71.1%
Chess move	29.3%	45.2%

(Du et al., 2023)

Recommended reading:

- Communicative Agents for Software Development, Qian et al., (2023)
- AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, Wu et al. (2023)

Andrew Ng

<https://youtu.be/sal78ACtGTc>



You can customize

- The finetuned chat LLM
- The embedding model
- The vector collection for the knowledge base
- System prompts

**Demo: Change to another model
and another knowledge base**

Models 12,139

gguf

new Full-text search

Sort: Trending

NousResearch/Hermes-2-Pro-Llama-3-8B-GGUF

Updated 4 days ago • 12.1k • 99

microsoft/Phi-3-mini-4k-instruct-gguf

Text Generation • Updated 15 days ago • 96.6k • 293

bartowski/Llama-3-ChatQA-1.5-8B-GGUF

Text Generation • Updated 5 days ago • 2.87k • 31

shenzhi-wang/Llama3-8B-Chinese-Chat-GGUF-8bit

Text Generation • Updated about 12 hours ago • 38.5k • 82

crusoeai/Llama-3-8B-Instruct-Gradient-1048k-GGUF

Updated 3 days ago • 11.3k • 57

xtuner/llava-llama-3-8b-v1_1-gguf

Image-to-Text • Updated 8 days ago • 8.92k • 52

NeverSleep/Llama-3-Lumimaid-8B-v0.1-GGUF

Updated about 8 hours ago • 2.74k • 23

NexaAIDev/octopus-v4-gguf

Updated about 14 hours ago • 683 • 22

xtuner/llava-phi-3-mini-gguf

Image-to-Text • Updated 8 days ago • 9.52k • 57

bartowski/Meta-Llama-3-70B-Instruct-GGUF

Text Generation • Updated 5 days ago • 6.05k • 20

PrunaAI/Llama-3-8B-Instruct-Gradient-1048k-GGUF-sma...

Updated 3 days ago • 16.9k • 28

QuantFactory/Meta-Llama-3-8B-Instruct-GGUF

Text Generation • Updated 17 days ago • 179k • 206

alexcovo/Meta-Llama-3-120B-Instruct-Q4_K_M-GGUF

Updated 2 days ago • 18

cognitivecomputations/Meta-Llama-3-120B-Instruct-gg...

Updated 1 day ago • 441 • 18

Orenguteng/Llama-3-8B-Lexi-Uncensored-GGUF

Updated 14 days ago • 19.8k • 40

Lewdicolous/Poppy_Porpoise-v0.7-L3-8B-GGUF-IQ-Imatr...

Updated about 22 hours ago • 14.1k • 24

bartowski/Llama-3-8B-Instruct-Coder-GGUF

Text Generation • Updated 2 days ago • 4.62k • 14

lmstudio-community/Meta-Llama-3-70B-Instruct-GGUF

Text Generation • Updated 4 days ago • 52.2k • 90

Lewdicolous/Poppy_Porpoise-0.72-L3-8B-GGUF-IQ-Imatr...

Updated 3 days ago • 2.14k • 13

Lewdicolous/Llama-3-Soliloquy-8B-v2-GGUF-IQ-Imatrix

Updated about 17 hours ago • 2.17k • 13



Bojan Tunguz ✓ @tunguz · Apr 21
AGI will be built with Python.

Let that sink in.

519 382 5,084 3.8M



Elon Musk ✓ @elonmusk

Rust

4:40 AM · Apr 22, 2023 · 3.7M Views

682 Retweets 333 Quotes 10.4K Likes 334 Bookmarks



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.



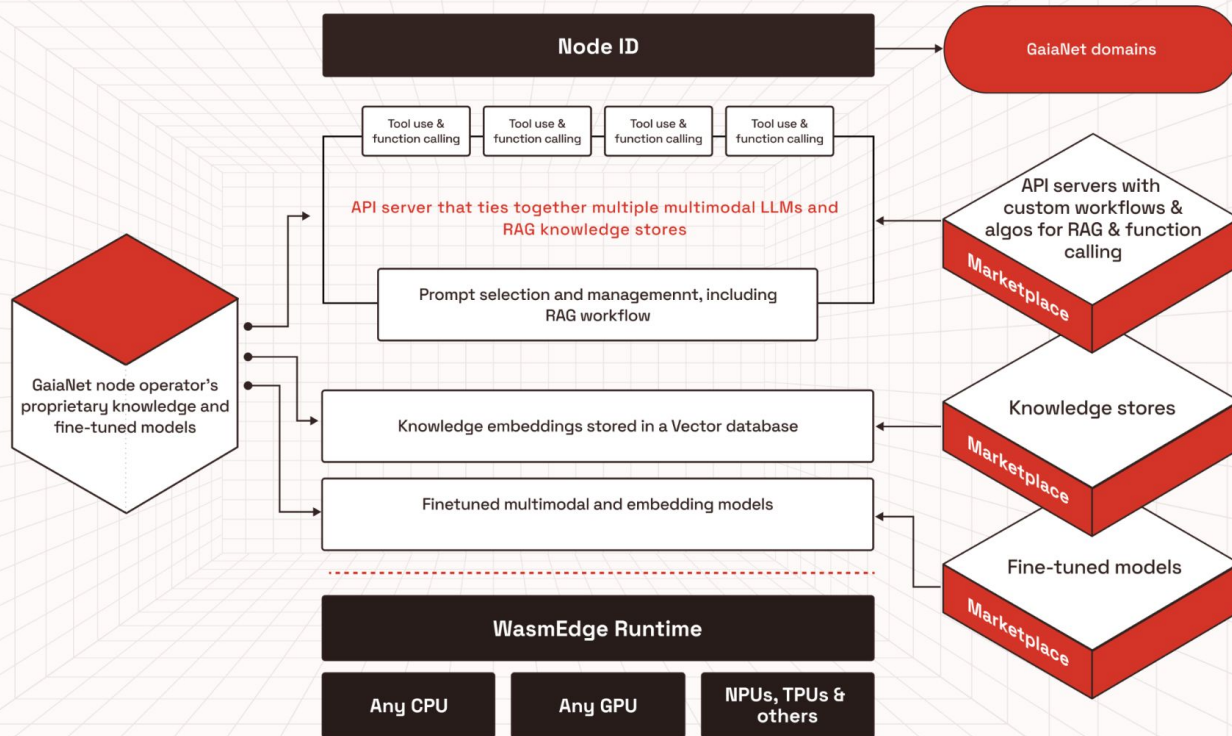
Chris Albon @chrisalbon · 1d

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

<https://blog.stackademic.com/why-did-elon-musk-say-that-rust-is-the-language-of-agi-eb36303ce341>

</> TECH STACK





Wasm-based runtime

- Very lightweight and fast
 - Entire runtime + app is less than 100MB
 - Runs well on Raspberry Pi and Jetson devices
 - Full native GPU and hardware accelerator support
- Single command to install and run as an unprivileged user – no daemon, no sudo
- Supports a wide range of LLMs, VLMs, MoE models on Hugging face out of the box
- 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
- Customizable formatted responses (JSON and function calling)

```
docker pull pytorch/pytorch:2.3.0-cuda  
12.1-cudnn8-runtime
```

[Copy](#)

Digest	OS/ARCH	Compressed Size ⓘ
0279f7aa2997	linux/amd64	3.47 GB

```
docker pull pytorch/pytorch:2.3.0-cuda  
11.8-cudnn8-devel
```

[Copy](#)

Digest	OS/ARCH	Compressed Size ⓘ
e0a9d9942dca	linux/amd64	8.73 GB



The vector database

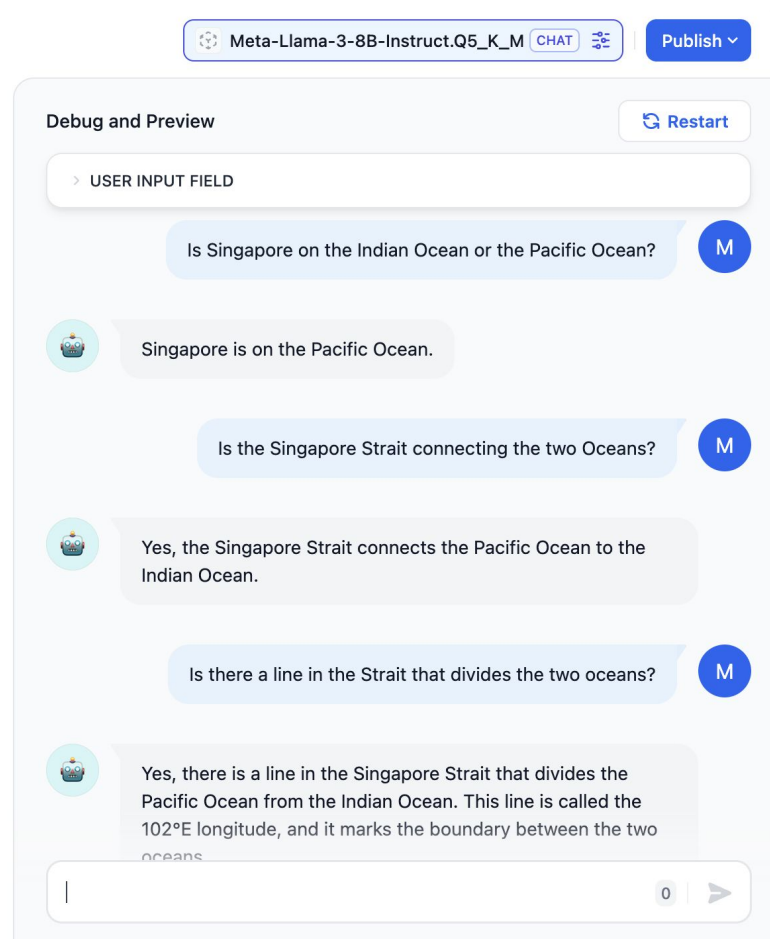
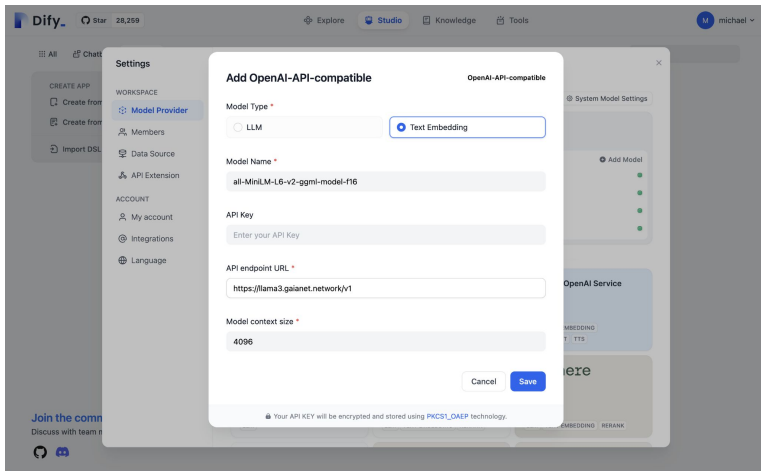
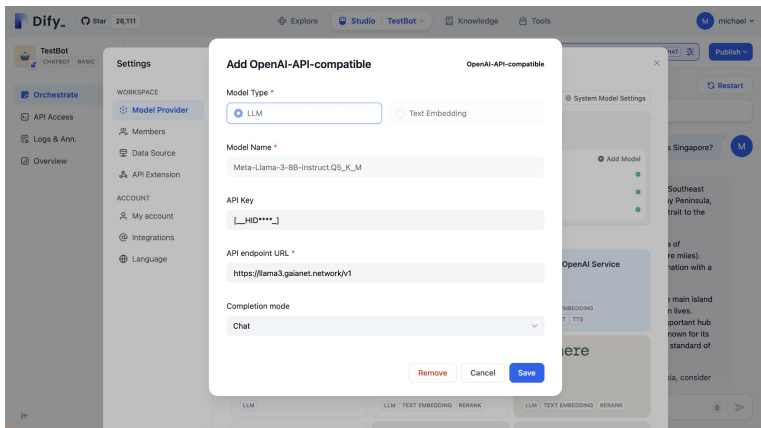
- Qdrant
 - Implemented in Rust
 - High performance
 - Feature rich
 - Scalable
 - Lightweight
- Others are also being supported
 - sqlite-vss
 - pinecone
 - etc

GaiaNet is a developer platform

- Use PyTorch / llama.cpp to finetune
- Use LangChain / LlamaIndex to create the knowledge base or vector collection
- **Use GaiaNet to run the service!**



Demo: OpenAI compatible service



<https://docs.gaiainet.ai/user-guide/apps/dify>



Public nodes for API services

Config option	Value
API endpoint URL	https://vitalik.gaianet.network/v1
Model Name (for LLM)	vitalik.eth-7b-q5_k_m
Model Name (for Text embedding)	all-MiniLM-L6-v2-ggml-model-f16
API key	Empty or any value


Config option	Value
API endpoint URL	https://0xc5f90fa1812dd7e27a26f1699954fe2d72e72299.gaianet.network/v1
Model Name (for LLM)	Meta-Llama-3-8B-Instruct-Q5_K_M
Model Name (for Text embedding)	nomic-embed-text-v1.5-f16
API key	Empty or any value

Config option	Value
API endpoint URL	https://llama3.gaianet.network/v1
Model Name (for LLM)	Meta-Llama-3-8B-Instruct.Q5_K_M
Model Name (for Text embedding)	all-MiniLM-L6-v2-ggml-model-f16
API key	Empty or any value

Config option	Value
API endpoint URL	https://0xf8bf989ce672acd284309bbbf4debe95975ea77.gaianet.network/v1
Model Name (for LLM)	Meta-Llama-3-70B-Instruct-Q5_K_M
Model Name (for Text embedding)	all-MiniLM-L6-v2-ggml-model-f16
API key	Empty or any value

<https://docs.gaianet.ai/user-guide/nodes>

Agent apps

 E2B

E2B_DEV
@E2B_DEV

AI Agents Landscape

By E2B.dev - Cloud Runtime for AI Agents

E2B users or integrations
INTEGRATED WITH E2B

Open sourceClosed Source

Coding

Open Interpreter
E2B INTEGRATION

Maige
RUNNING ON E2B

Sweep AI

WorkGPT

Vanna.AI

DemoGPT

AutoPR

Aide

Smol Developer

bloop.

Automata

Continue

GPT Migrate

GPT Engineer

CodeFuse

Stackwise

Sourcegraph Cody AI

cody

ReactAgent

GPT Pilot

English Compiler

Productivity
+ Daily Life

Local GPT

Allie

PromethAI

Agent4Rec

General Purpose

Promptly

AutoGPT

BeeBot

ChatArena

BabyAGI

Multiagent Debate

GPTDiscord

evo.ninja

MiniAGI

MultiGPT

XAgent

Web3 GPT

Suspicion Agent

Tusk

BitBuilder

v0 by Vercel

autopilot

phind

Alpine Autopilot

Factory

Deepnote AI

Copilot X

HEX Hex Magic

codium

GitLab Duo

GitWit
REACTEVAL
RUNNING ON E2B

MakeDraft

Dosu

CodeWP

grit

Input

Kusho

SECOND

mutable.ai

Butternut AI

Cursor

Codegen

Duckie AI

DevGPT

Moone

MultiOn

Lindy

Spell

Claros AI Shopper

iMean.AI

AgentScale

Cykel

FL DE

Otherside Personal Assistant

Wispy

ollie

COGNOSYS

Raycast

Chathelp

Lutra

Artisian

Sentius

magic loops

B2 AI

GOD MODE

ADEPT

ChatGPT for Slack

Questflow

HR

Autonomous HR Chatbot

Data Analysis

LangChain
E2B Data Analyst
INTEGRATED WITH E2B

MemGPT

TalktoData

Julius

AskYourDatabase

Dot

Graphlit

Business Intelligence

Kompas AI
RUNNING ON E2B

Taxy AI

clay

ability.ai

Juno

aomni

Science

Chem Crow

NLSOM

Research

GPT Researcher

Design

Diagram

Marketing

Blobr

GoCharlie

AskToSell

Build Your Own

Superagent
INTEGRATED WITH E2B

CHATDEV

BondAI

Adala

LLMStack

AgentPilot

AgentGPT

Agents

IX

AutoGen

MetaGPT

pezzo

AgentVerse

AgentForge

OpenAgents

SuperAGI

LangChain

Agents Runtime

E2B

Google Cloud Platform

aws

Modal

Azure

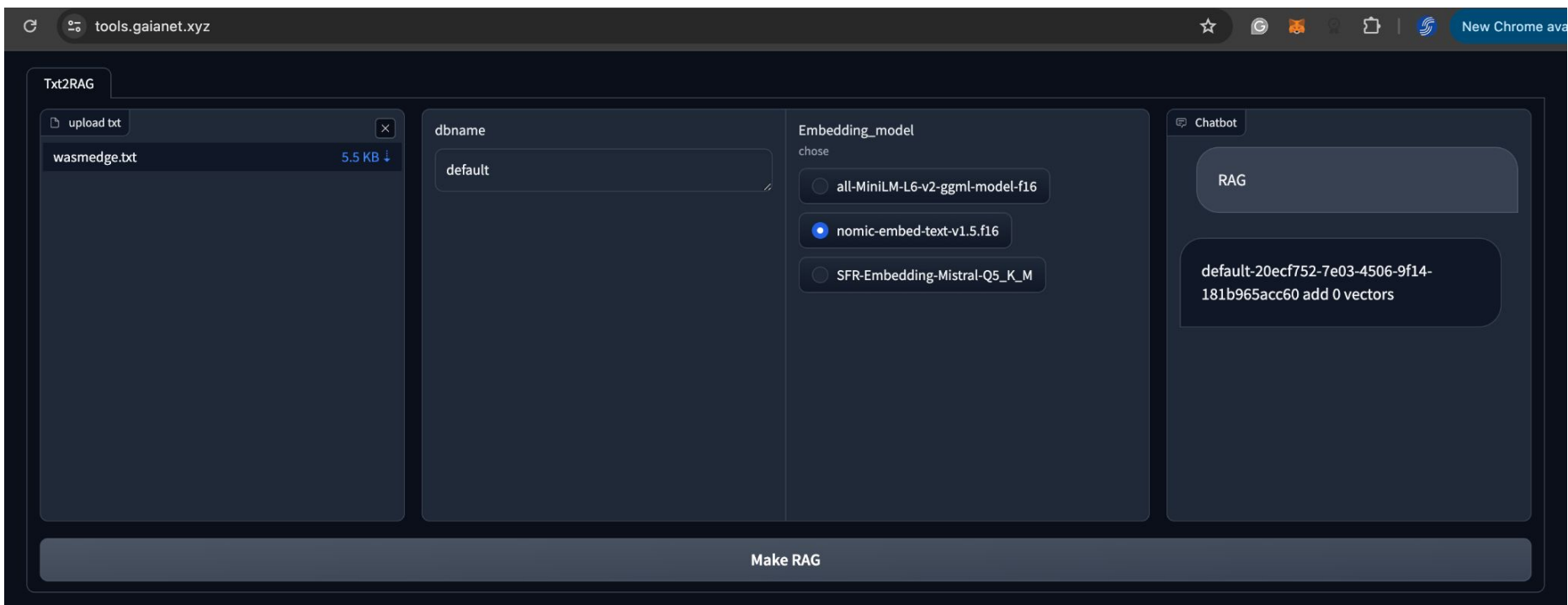
MADE BY E2B

APR 2024 * V2.4

LLM frameworks



Demo: Create a customized knowledge base



<https://tools.gaianet.xyz/>



More ways to create a knowledge base

- Text paragraphs: <https://docs.gaianet.ai/creator-guide/knowledge/text>
- Markdown: <https://docs.gaianet.ai/creator-guide/knowledge/markdown>

Try this node for a high school chemistry teacher:

<https://0xc5f90fa1812dd7e27a26f1699954fe2d72e72299.gaianet.network/>



Next steps

- Multimodal input
 - Early access: <https://www.secondstate.io/articles/llava-v1.6-vicuna-7b/>
- Multimodal output
- Function calling
- Incentivized nodes

<https://gaianet.ai/>



Home

Network map

Domain

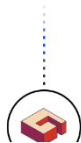
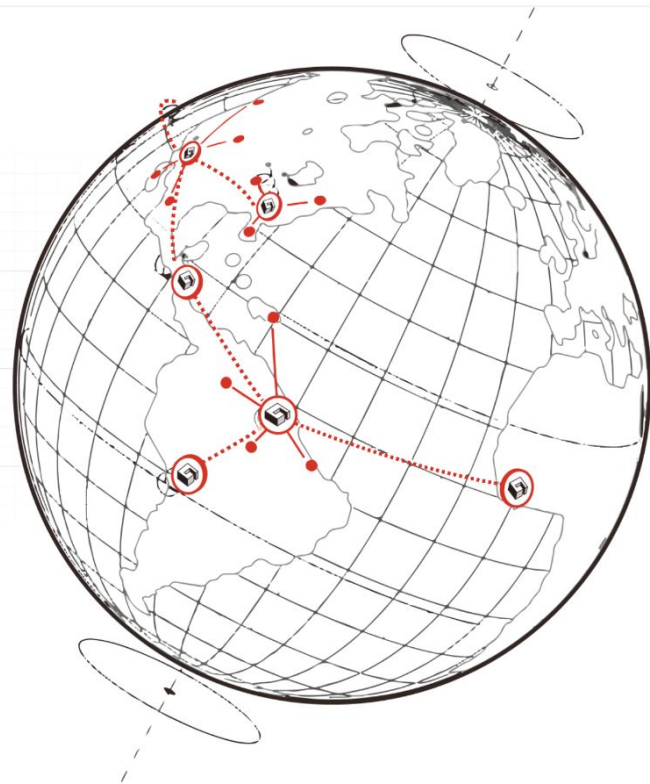
Install Node

Docs

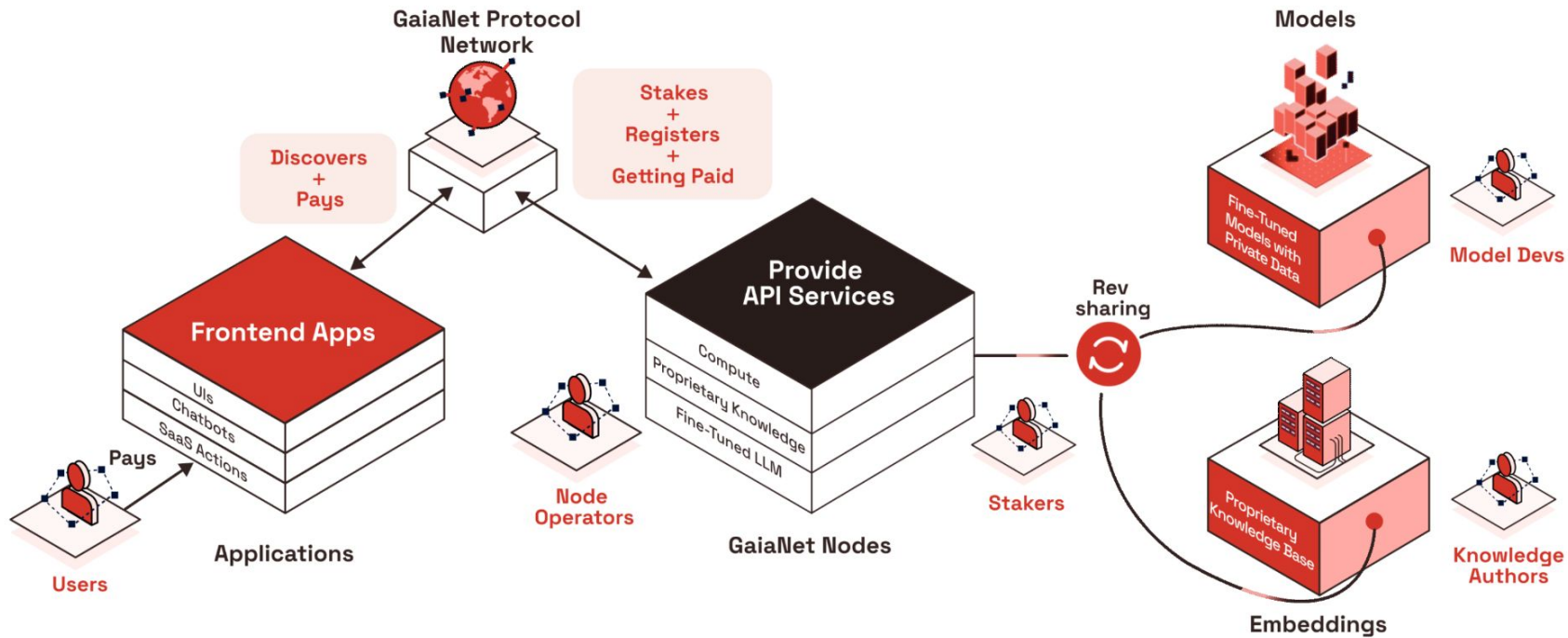
Connect ↗

Decentralized GenAI Agents Network.

INSTALL NODE



HOW IT WORKS



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>

Thank you

Learn more:

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

