AI powered Rust programming and LLM Agents

Miley Fu-WasmEdge

CNCF Ambassador

KubeCon+Open Source Summit+AI_Dev China 24 Co-Chair

https://github.com/WasmEdge/WasmEdge

Content

RustCoder

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 Víquez
@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

RustCoder: A coding assistant https://flows.network/learn-rust

💻 RustCoder: Your AI Rust Programming Assistant

RustCoder is designed to help you learn and master Rust programming through intelligent assistance and guidance. Our AI assistant understands Rust's unique features and helps you write efficient, safe code.

📚 Backed by authoritative Rust content
The Learn Rust assistant answers your questions based on official documentation and tutorial content from the Rust Foundation. It provides high quality explanations with code examples using a large language model's own internal knowledge of computer programming.

🔄 Interactive learning
Built on conversation capabilities of ChatGPT/4, the Learn Rust assistant can guide you to the right answers through back and forth QA. You can ask for clarifications, further explanations, and additional code examples.

🌐 Accessible to diverse learners
Most Rust learning materials and docs are English only. But the Learn Rust assistant can frequently converse in almost all major languages, such as Chinese, Japanese, Korean, Hindu etc. It bridges the gap between English content authors and underserved language communities.

⏰ Always available
The Learn Rust assistant is your dedicated learning companion, available anytime anywhere. It is always there when you have questions or need a fresh set of eyes on your code.

RustCoder: AI 助力 Rust 学习

- Rust 编码助手，可以解释 Rust 概念、编写 Rust 代码、完成 Rust 算法并修复错误。
- 兼容 OpenAI API：可以与任何流行的 AI IDE 集成，如 bolt.new、cursor、Zed 和 Continue
- Powered by Open Source
  - 知识：书籍、Rust 示例和数据结构与算法（Rust 语言）
  - 运行时：WasmEdge
  - 模型：QwenCoder
  - Host：Gaia

Config option	Value
API endpoint URL	https://rustcoder.gaia.domains/v1
Model Name (for LLM)	rustcoder
Model Name (for Text embedding)	nomic-embed
API key	Empty or any value

效果展示: 开放原子大赛与开源操作系统训练营联合推出的Rust数据结构与算法学习赛

题目1：自己设计实现一个统计不重复元素个数的算法，输入为逗号分隔的字符串。（20）

输入：1,爱,好,0,100,爱,1,物理,化学,物理,AI,AI
输出：8

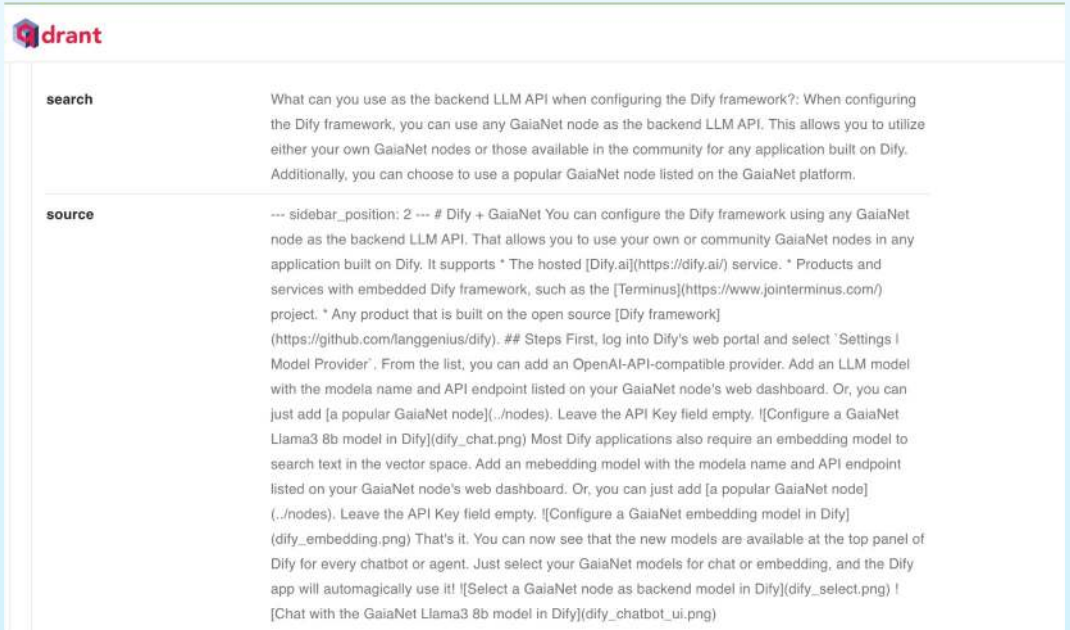
效果展示

```
题目3: 实现一个 Rust 算法，输入是人数(>=2)，计算任意一天同时存在两个及以上的人过生日的概率，保留四位小数。

输入: 50
输出: 0.9704
输入: 61
输出: 0.9951
```

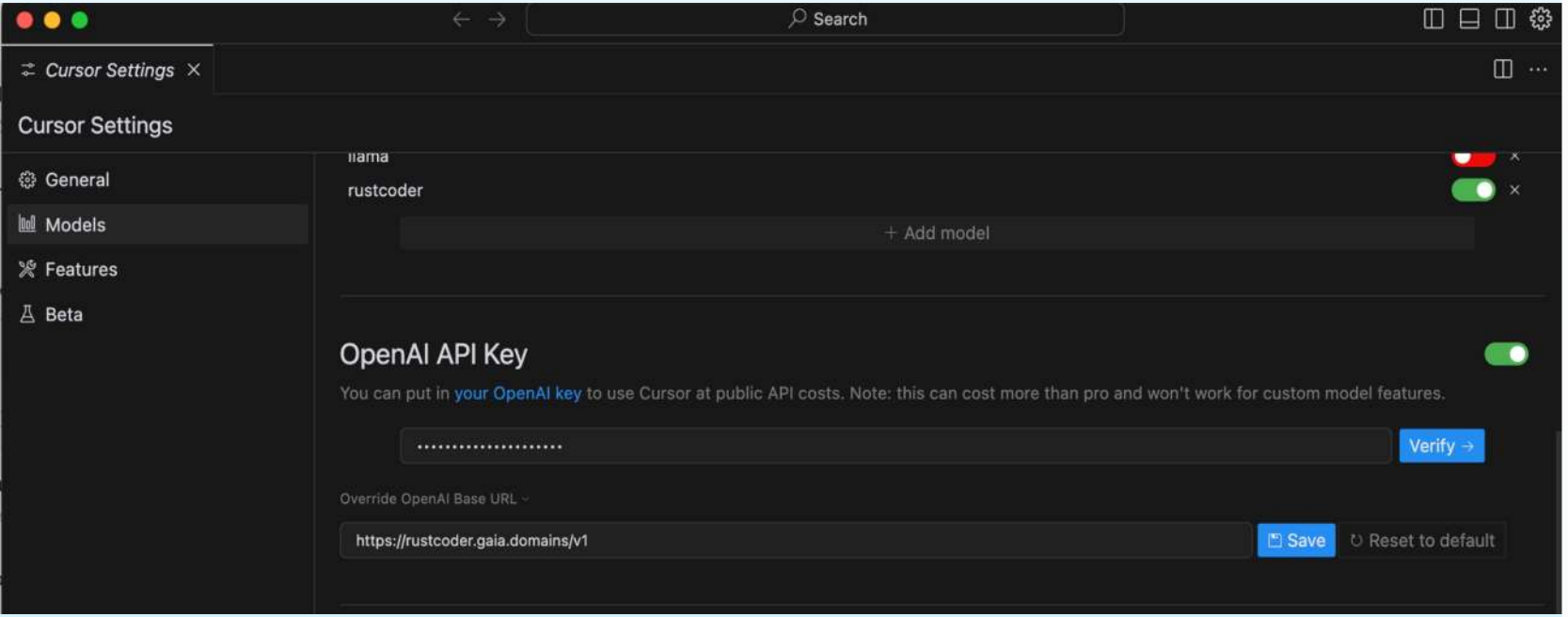
这是如何做到的

- 许多开源模型现在支持长达 128k token 的长上下文长度，例如 llama-3.2-3B 和 Qwen-2.5-coder
- 更长的上下文长度可以处理大量文本，但也有弊端：
  - 推理时间增加。
  - 计算资源要求更高。
- 我们的解决方案：
  - 根据用户问题选择并提供最相关的章节或部分作为上下文。
  - 优化效率，同时保持高质量的响应。



如何使用 RustCoder

通过与 OpenAI 完全兼容的 API Point
与 AI IDE 相结合，比如 Cursor 和 Zed



Open recent project

untitled | settings.json ~/.config/zed

~/.config/zed/settings.json

```
{
  "features": {
    "inline_completion_provider": "none"
  },
  "language_models": {
    "openai": {
      "version": "1",
      "api_url": "https://rustcoder.gaia.domains/v1",
      "low_speed_timeout_in_seconds": 60,
      "available_models": [
        {
          "name": "rustcoder",
          "max_tokens": 8096
        }
      ]
    }
  },
  "assistant": {
    "provider": "openai",
    "default_model": {
      "provider": "openai",
      "model": "rustcoder"
    },
    "version": "2"
  },
```

Cursor + RustCoder Demo

在自己本地运行

A Qwen2.5-Coder-7B node with Rust language docs

Step 1: Install GaiaNet node

```
curl -sSfL 'https://github.com/GaiaNet-AI/gaianet-node/releases/latest/download/install.sh' | bash
```

Step 2: Init with the Qwen2.5 Coder 7B model and Rust books

```
gaianet init --config https://raw.githubusercontent.com/GaiaNet-AI/node-configs/main/qwen-2.5-coder-7b-instruct_rust
```

Step 3: Start the node

```
gaianet start
```

Now you can use the node as a web-based chatbot or as an OpenAI API drop-in replacement.

清华 Rust 训练营

KCD GROWING CLOUD NATIVE TOGETHER BEIJING

训练营小助手 Rustcoder

2025/01/17 17:31:27

国内外已有众多高校在大量使用rust，清华大学的新生学习用的操作系统是什么？

2025/01/17 17:31:27

根据描述，清华大学新生学习用的操作系统是rCore和zCore。

复制

Learning Rust Community

Camps

Camp List

Rust Beginners Learning Camp (...

Stage 0: Pre-class

Stage 1: Basics

Stage 2: Professional

Stage 3: Projects

Rust Beginners Learning Camp (Data Structures and Algorithms)

Date：2025/01/17-2025/03/09

Organized by：Rust 基金会
RustCC 社区
SecondState
清华大学开源操作系统训练营 联合主办

Growth Path

Preliminary understanding of the training camp

第一期 Rust 入门训练营（数据结构与算法）

未来计划

- 将 RustCoder 与 Rust Playground 集成
  - 利用 Rust 编译器强大的错误检测功能
  - 使用错误消息和 RustCoder 高效地调试生成的 Rust 代码。
- 支持更多的 Rust 知识库

Translation Agent for English Workshops/ Tutorials

VideoLangua.com: High quality translated videos

Translate any video to your language

Upload any video file and let us translate it for you! You can choose to dub (voice-over) the video with the translated language, or add closed-captions while preserving the original soundtrack. Currently available in

English to Chinese

English to Japanese

Short videos are free

Videos under 3 minutes will be translated for free! Share them in your social channels!

Voice-over or closed-caption

The translation could be added to the original video as a dubbed soundtrack or closed-captions.

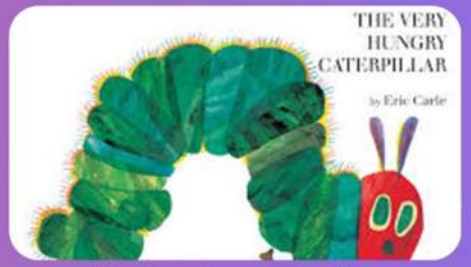
AI powered

The translation service is powered by state-of-the-art AI models running in the Gaia Network.

Examples

Geoffrey Hinton On working with Ilya choosing problems and the power of intuition

Linus's opinion about AI coding



The Very Hungry Caterpillar

Powered by WasmEdge and Gaia

videolangua.com

Translate any video to your language

Upload any video file and let us translate it for you! You can choose to dub (voice-over) the video with the translated language, or add closed-captions while preserving the original soundtrack. Currently available in

English to Chinese | English to Japanese | English to korean

Chinese to English | Japanese to English

Short videos are free

Videos under 3 minutes will be translated for free! Share them in your social channels!

Voice-over or closed-caption

The translation could be added to the original video as a dubbed soundtrack or closed-captions.

AI powered

The translation service is powered by state-of-the-art AI models running in the Gaia Network.

Feedback

Examples

Geoffrey Hinton On working with Ilya choosing problems and the power of intuition

Linus's opinion about AI coding

The Very Hungry Caterpillar

Brought to you by

AI models on the edge

KCD
GROWING CLOUD NATIVE TOGETHER
BEIJING

Whisper

—

Transcribe text with timestamps. Supports 90 languages.

LLM

—

Multiple LLMs to clean up transcribed text, translate, and check translation quality.

TTS

—

Use user-defined LORA to generate voice that is natural sounding for specific domains.

Takeaways

- Supports multiple GenAI models
- Lightweight (<20M of total application size)
- Portability across GPU platforms

Takeaway – scale the inference

- Requires multiple specialized models
- Requires app to be tightly coupled with the model
- Requires efficient use of GPU compute

Both agents built on LlamaEdge

LlamaEdge: a universal GenAI runtime

- OpenAI compatible API server
  - Support multimodal APIs
  - Support tool calls
  - Support built-in search
  - Support RAG (OpenAI assistant API)
- A component library for developers to roll-your-own tightly coupled LLM apps
  - Rust
  - JavaScript
- Based on Linux Foundation’s WasmEdge project

https://github.com/LlamaEdge/LlamaEdge

Why not Ollama?

Model selection

Ollama only supports LLMs and recently VLMs.

But agent apps need to incorporate many models, including traditional vision, audio, and OCR models based on Torch.

Operational weight

Ollama is a large GO app built on top of llama.cpp. It incorporate multiple platform-specific binaries for portability.

It requires a sudo daemon and access to its proprietary model hub.

Why not llama.cpp, whisper.cpp, vLLM etc?

Portability

Native inference apps developed and tested on a **Macbook** cannot run on a Linux server with AMD or Nvidia or ARM or Huawei accelerators.

Modern tools

Difficult to use C/C++, CUDA, ROCm, CANN, metal / MLX functions in modern web app frameworks.

Safety

Prone to memory errors and crashes. Requires containers or VMs to isolate from the OS.

LlamaEdge is a developer platform

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

Run open source GenAI models
on your infra

and / or

bundled with your own apps

Key features

- Lightweight
  - Less than 50MB
  - No Python dependency
- Portable
  - Develop on one GPU and deploy on another without recompiling or code changes
  - Support mainstream GPUs and NPUs out of the box
- Wide selection of AI models
- Embeddable
- Cloud-native and supported in major distros

Deploy production-ready LLM apps on LlamaEdge

- Build a single **portable** and deployable app
  - Move code closer to model and data
  - Improve efficiency
  - Simplify development and workflow
  - Improve security
- No need for external middleware and containers to orchestrate common LLM app components
- 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

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!

Wikifunctions, a serverless function platform integrated into Wikipedia, one of the world's most popular websites

over 120k views and 400+ upvotes in just half a day.

r/programming · 14 days ago
RustyLanguage

Wikimedia Slashed 300ms Off Every WASM Execution with WasmEdge

wikifunctions.org

Open

650 | 119 | Share

Post Insights

Only the post author and moderators can see this

238K	96%	119	149
Total Views	Upvote Rate	Comments	Total Shares

Hourly views for first 48 hours

Officially supported by

Calling for contributors

LFX Mentorship

Google Summer of Code

is:issue state:open

Open 203 Closed 1,299 Author Labels Projects Milestones

- **LFX Workspace: Create a Japanese translation agent for CNCF videos** LFX Mentorship
  #4047 · ainozaki opened last week
- **question: how to dump an executable bitcode and LLVM IR** question
  #4046 · hly2019 opened 2 weeks ago
- **LFX Workspace: Component Model's Validator** LFX Mentorship
  #4044 · sridamul opened 2 weeks ago
- **LFX Workspace: Implement a new WasmEdge installer in Rust** c-Installer LFX Mentorship
  #4039 · alan910127 opened 2 weeks ago
- **LFX Workspace: Rust Coder** LFX Mentorship
  #4038 · Acuspeedster opened 2 weeks ago
- **feat: Support HIP backend for the WASI-NN llama.cpp plugin on AMD platforms.** enhancement WASI-NN
  #4037 · hydai opened 2 weeks ago

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

WeChat

YouTube

WasmEdge Github

Stay in Touch!

Github/X/ LinkedIn

@mileyfu

GOSIM®
GLOBAL OPEN-SOURCE INNOVATION MEETUP

GOSIM AI Paris 2025

The open source AI conference

May 6–7, 2025 Station F Hours: 9:00 am – 6:00 pm

Tracks

- AI Model
- AI Infra
- AI Apps
- Embodied AI
- AI for Science
- PyTorch Day

Speakers

- PyTorch
- Dify 82kstars
- AutoGen 41.4kstars
- GLM 41kstars
- TiDB 38.1Kstars
- OpenManus 33.6kstars
- MiniCPM
- Open-R1
- Jina.ai 21kstars
- Qwen 17.4kstars
- MCP 17kstars
- SGlang 11.9Kstars
- OWL 11.5kstars
- FlagAI 3.9kstars

KCD
GROWING CLOUD NATIVE TOGETHER
BEIJING

KubeCon | CloudNativeCon
China 2025

10-11 JUNE
HONG KONG
#KUBECON
#CLOUDNATIVECON

REGISTER
SPONSOR
VIEW THE SCHEDULE

KCD
GROWING CLOUD NATIVE TOGETHER
BEIJING

Thanks