
Kubernetes Community Days Shanghai 2024
KCD
GROWING CLOUD NATIVE TOGETHER
SHANGHAI
边缘上的可移植、高性能的 LLM 推理
夏歌
X：@alabulei
WasmEdge Runtime: https://github.com/WasmEdge/WasmEdge

We love Open Source

- WasmEdge：轻量级+跨平台的 AI runtime
  - https://github.com/WasmEdge/WasmEdge
- LlamaEdge：基于 WasmEdge，是LLM 应用的开发者平台
  - https://github.com/LlamaEdge/LlamaEdge

Content

目录

01 在你的设备上运行 Llama-3-8B

02 开源 VS 闭源：为什么不能只用 OpenAI

03 使用 LlamaEdge 构建 RAG 应用

04 未来规划

Part 01

在你的设备上运行 Llama-3-8B

一键运行 Llama-3-8B

```
bash <(curl -sSfL 'https://raw.githubusercontent.com/LlamaEdge/LlamaEdge/main/run-llm.sh') --model
llama-3-8b-instruct
```

Demo 视频请查看：
https://www.bilibili.com/video/BV1Tr42137Pu

base ~

一键运行 Llama-3-8B

- 下载 LLM runtime： WasmEdge
- 下载 Llama-3-8B GGUF 格式的模型
- 下载 LLM inference 的 Wasm 文件
- 下载 chatbot ui
- 运行 Llama-3-8B

Part 02

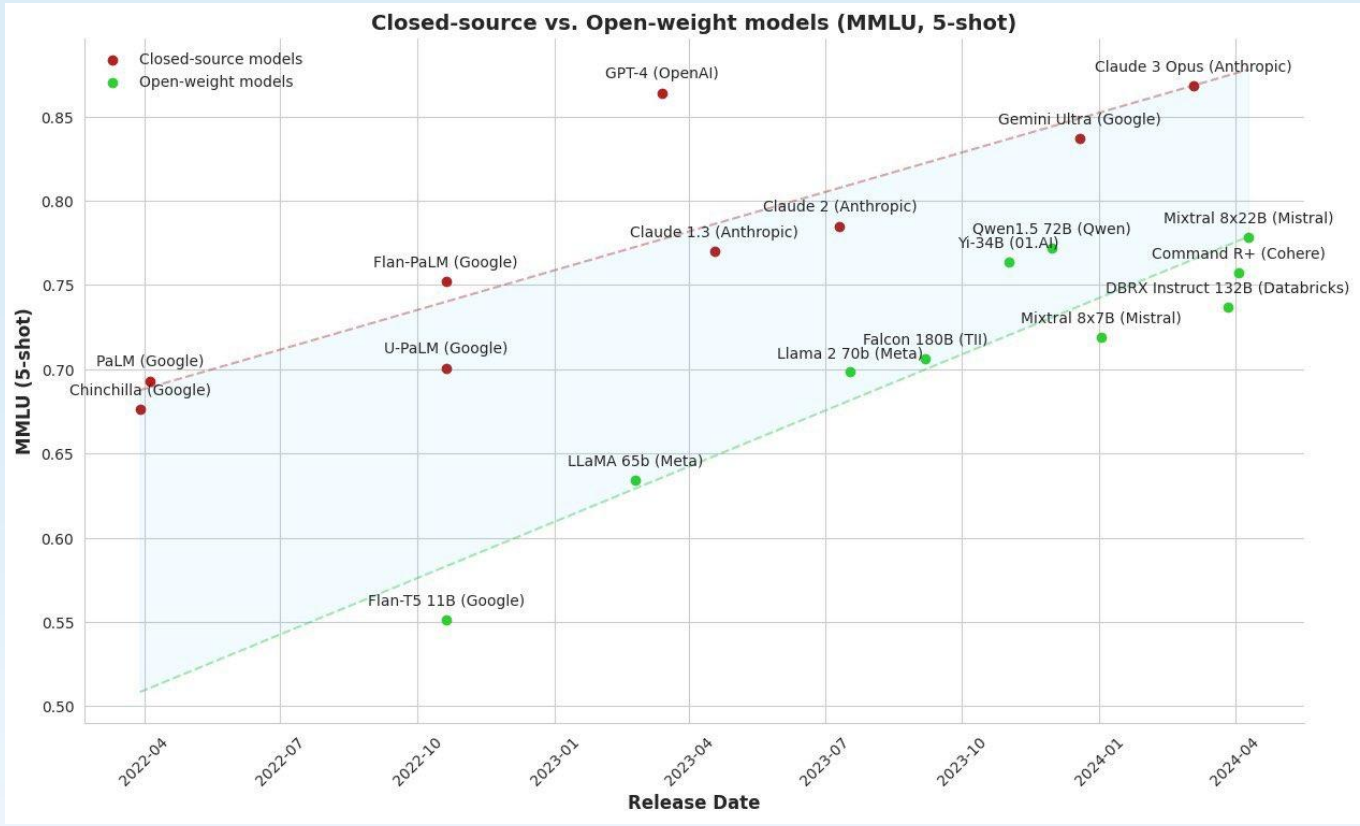
开源 VS 闭源

为什么不能只用 OpenAI

为什么不能只用 OpenAI

- 改造 ChatGPT 很难
- 没有隐私与控制
- 审查与偏见
- 贵！


开源模型会越来越落后



Jim Fan 
@DrJimFan

Follow

The upcoming Llama-3-400B+ will mark the watershed moment that the community gains open-weight access to a GPT-4-class model. It will change the calculus for many research efforts and grassroot startups. I pulled the numbers on Claude 3 Opus, GPT-4-2024-04-09, and Gemini.

Llama-3-400B is still training and will hopefully get even better in the next few months. There is so many research potential that can be unlocked with such a powerful backbone. Expecting a surge in builder energy across the ecosystem!

	A	B	C	D	E	F
1	**Benchmark**	**Llama-3-400B+**	**Claude-3-Opus**	**GPT-4-turbo**	**Gemini Ultra 1.0**	**Gemini Pro 1.5**
2	MMLU	86.1	86.8	86.5	83.7	81.9
3	GPQA	48	50.4	49.1	-	-
4	HumanEval	84.1	84.9	87.6	74.4	71.9
5	MATH	57.8	60.1	72.2	53.2	58.5
6						

为什么要用 LlamaEdge API server

- 轻量且快速
  - 整个运行时 + 应用程序小于 30MB
  - 可以在 Raspberry Pi 和 Jetson 设备上运行
  - 完整的原生 GPU 和硬件加速器支持
- 无需以超级用户的身份安装和运行
- 可以通过容器工具和 k8s 直接管理和编排
- 开箱即用，支持 Hugging Face 上的任何 GGUF 模型
- 支持广泛的设备和驱动程序，可以以本机 GPU 速度运行，如CUDA、TensorRT、Apple Metal
- 可定制的格式化响应（JSON 和函数调用）
- 高效且可扩展的开发者平台
  - RAG、会话状态和函数调用都可以像 OpenAI Assistant API 一样内置到 API 服务器中
  - 不需要单独的中间件应用程序（例如LangChain）

LlamaEdge 是一个开发者平台

- 构建单个可移植易部署的应用程序
  - 提高效率
  - 简化开发和工作流程
  - 提高安全性
- 无需外部中间件和容器来编排常见的 LLM 应用程序组件
- 没有 Python 依赖（例如 LangChain）
- 使用 Rust 扩展 LlamaEdge 组件!
- 与 OpenAI 最佳性能相匹配的开发体验
  - 高度集成的OpenAI Assistant API

Demo

- 在 mac 上，将 Rust 代码编译成 Wasm 文件
- 把这个 Wasm 复制到 NVDIA 设备上运行

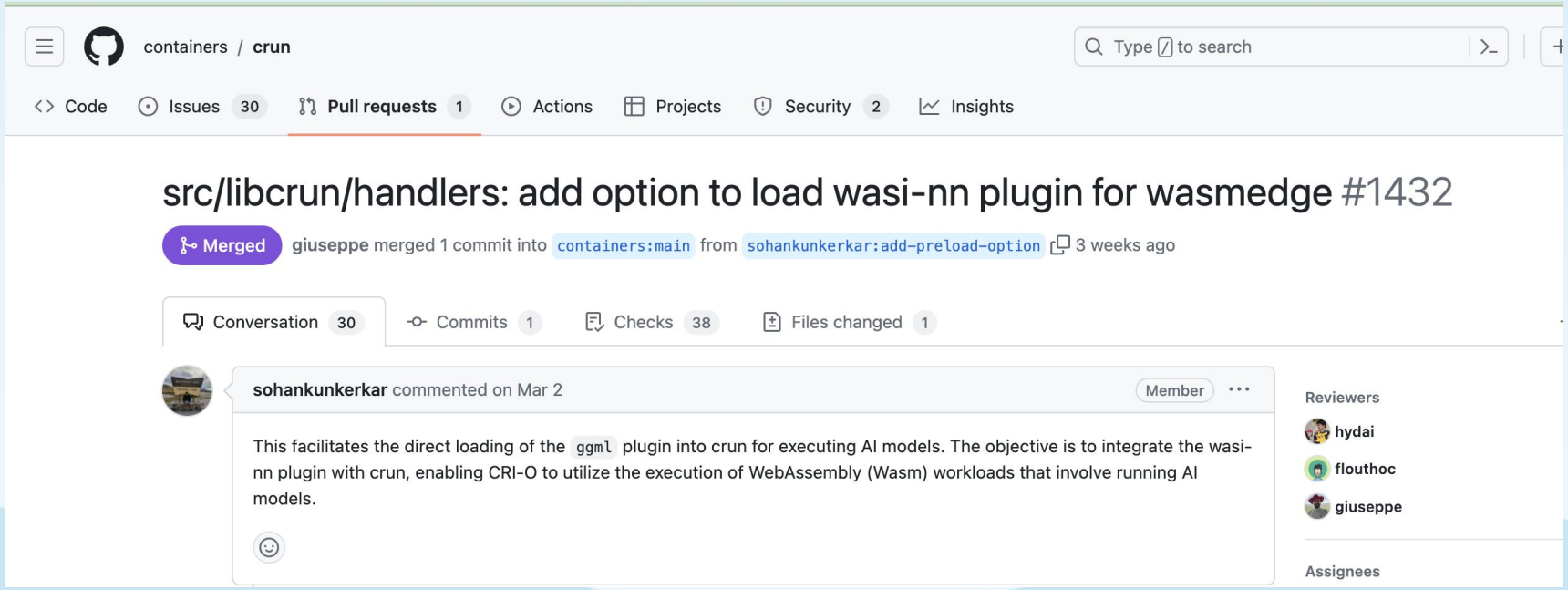
视频请查看：
https://www.bilibili.com/video/BV1Tr42137Pu

Dev

- 使用多种不同的语言来创建您的应用程序
  - 目前支持 Rust，但 JavaScript 将很快发布。
- 只需要调用 WasmEdge API 即可进行推理操作。
  - 无需担心 GPU 驱动程序或张量库。
  - WasmEdge 推理 API 基于 W3C 的 WASI NN 标准。
- 将应用程序编译为 Wasm。
- 使用现有工具分发和部署 Wasm 二进制文件。

Ops

- 安装 WasmEdge 以及 ggml plugin
  - 它将为此设备安装 GPU 驱动程序和 SOTA 推理库
- 运行 Wasm 二进制应用程序
- Bonus：WasmEdge 运行时本身是一个安全沙箱，可以通过 K8s、Docker 和 OpenShift 等容器工具进行管理



Part 03

使用 LlamaEdge 构建 RAG 应用

构建 RAG 应用的流程

- 文本分段
- 使用向量数据库+开源 embedding model 向量分好段的文档并存储在向量数据库
- 当用户提问时，embedding 搜索相似向量，并返回给 Chat model
- Chat model 根据用户问题以及搜索到的向量回答用户问题

使用 LlamaEdge 构建 RAG 应用的方法

- 提供一个和 OpenAI 兼容的 API server，借助 Langchain构建客户端 RAG 应用
  - https://llamaedge.com/docs/user-guide/client-side-rag
- 直接构建一个兼容 OpenAI API 的 RAG API server
  - https://llamaedge.com/docs/user-guide/server-side-rag

使用 LlamaEdge 构建 RAG 应用

```
wasmedge --dir .:. \
   --nn-preload default:GGML:AUTO:Llama-2-7b-chat-hf-Q5_K_M.gguf \
   --nn-preload embedding:GGML:AUTO:all-MiniLM-L6-v2-ggml-model-f16.gguf \
   rag-api-server.wasm -p llama-2-chat --web-ui ./chatbot-ui \
     --model-name Llama-2-7b-chat-hf-Q5_K_M,all-MiniLM-L6-v2-ggml-model-f16 \
     --ctx-size 4096,384 \
     --rag-prompt "Use the following context to answer the question.\n----------------\n" \
     --log-prompts --log-stat
```

docker ~ +

```
base ~
docker run -p 6333:6333 -p 6334:6334 \
    -v $(pwd)/qdrant_storage:/qdrant/storage:z \
    -v $(pwd)/qdrant_snapshots:/qdrant/snapshots:z \
    qdrant/qdrant

                 _                 _
  __ _  __| |_ __ __ _ _ __ | |_
 / _` |/ _` | '__/ _` | '_ \| __|
| (_| | (_| | | | (_| | | | | |_
 \__, |\__,_|_|  \__,_|_| |_|\__|
    |_|

Access web UI at http://0.0.0.0:6333/dashboard

2024-04-20T04:24:32.718156Z  INFO storage::content_manager::consensus::persistent: Loading raft state
from ./storage/raft_state.json
2024-04-20T04:24:32.725554Z  WARN storage::content_manager::toc: Collection config is not found in the
 collection directory: "./storage/collections/.DS_Store", skipping
2024-04-20T04:24:32.725952Z  INFO storage::content_manager::toc: Loading collection: default
2024-04-20T04:24:32.845947Z  INFO storage::content_manager::toc: Loading collection: ktx-smaller
2024-04-20T04:24:33.027851Z  INFO qdrant: Distributed mode disabled
2024-04-20T04:24:33.027864Z  INFO qdrant: Telemetry reporting enabled, id: d87f49eb-ef3b-43da-93da-78b
70c70f1fa
2024-04-20T04:24:33.028658Z  INFO qdrant::actix: TLS disabled for REST API
2024-04-20T04:24:33.028693Z  INFO qdrant::actix: Qdrant HTTP listening on 6333
2024-04-20T04:24:33.028698Z  INFO actix_server::builder: Starting 7 workers
2024-04-20T04:24:33.028703Z  INFO actix_server::server: Actix runtime found; starting in Actix runtime
2024-04-20T04:24:33.029467Z  INFO qdrant::tonic: Qdrant gRPC listening on 6334
2024-04-20T04:24:33.029475Z  INFO qdrant::tonic: TLS disabled for gRPC API
```

Demo 视频请查看：

https://www.bilibili.com/video/BV1Tr42137Pu

Part 04
未来规划

roadmap

- 支持更多类型的 model
- 支持开发者自由组合 RAG 逻辑
- 语言 SDK
- 支持 function calling

相关资源

https://github.com/LlamaEdge/LlamaEdge

https://llamaedge.com/docs/intro

https://github.com/GaiaNet-AI

Thanks.