

More

LocalAI

LocalAI is the free, Open Source OpenAI alternative. LocalAI act as a drop-in replacement REST API that's compatible with OpenAI API specifications for local inferencing. It allows you to run LLMs, generate images, audio (and not only) locally or on-prem with consumer grade hardware, supporting multiple model families and architectures.

- [Overview | LocalAI documentation](#)
- GitHub: <https://github.com/mudler/LocalAI>

OpenAI Proxy

Proxy Server to call 100+ LLMs in a unified interface & track spend, set budgets per virtual key/user

Features:

- **Unified Interface:** Calling 100+ LLMs Huggingface/Bedrock/TogetherAI/etc. in the OpenAI ChatCompletions & Completions format
- **Cost tracking:** Authentication, Spend Tracking & Budgets Virtual Keys
- **Load Balancing:** between Multiple Models + Deployments of the same model - LiteLLM proxy can handle 1.5k+ requests/second during load tests.

```
“ ???? LLM
????????????????????????????????????????????????????????????
?? OpenAI Proxy ??????????????????
    • ?? API ????????
    • ???
    • ???
```

- Doc: https://docs.litellm.ai/docs/simple_proxy

Xinference

Xorbits Inference (Xinference) is an open-source platform to streamline the operation and integration of a wide array of AI models. With Xinference, you're empowered to run inference

using any open-source LLMs, embedding models, and multimodal models either in the cloud or on your own premises, and create robust AI-driven applications.

- [Welcome to Xinference! — Xinference](#)
- GitHub: <https://github.com/xorbitsai/inference>

NVIDIA NIM

Explore the latest community-built AI models with an API optimized and accelerated by NVIDIA, then deploy anywhere with NVIDIA NIM inference microservices.

- [NVIDIA NIM for Deploying Generative AI | NVIDIA](#)
- Doc: [Introduction - NVIDIA Docs](#)
- Models: [google / gemma-7b](#)
- YT: [Self-Host and Deploy Local LLAMA-3 with NIMs - YouTube](#)

text-generation-webui

A Gradio web UI for Large Language Models.

???????????????? API?

???????? AI ??

- Chat
- Fine-Tune Model
- Multiple model backends: Transformers, llama.cpp (through llama-cpp-python), ExLlamaV2, AutoGPTQ, AutoAWQ, GPTQ-for-LLaMa, QuIP#.
- OpenAI-compatible API server with Chat and Completions endpoints

??

- GitHub: <https://github.com/oobabooga/text-generation-webui>
- GitHub: <https://github.com/Atinoda/text-generation-webui-docker>
- [?????LLMs???? ???? \(?\) - HackMD](#)
 - YOUTUBE [[?? TextGen](#)]
 - YOUTUBE [[????????](#)]
 - YOUTUBE [[??AI??](#)]
 - YOUTUBE [[????](#)]
 - YOUTUBE [[??????](#)]

- ??? [Z01_TextGen_Colab.ipynb](#)
- ?????????? (account:nchc password:nchc) ?????

Revision #4

Created 11 November 2024 09:39:31 by Admin

Updated 11 November 2024 09:49:52 by Admin