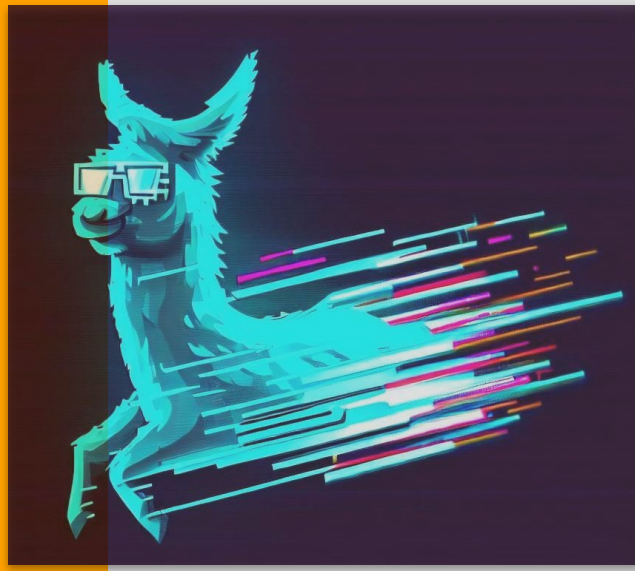
LocalAI:
The free, Open Source OpenAI alternative.



$whoami

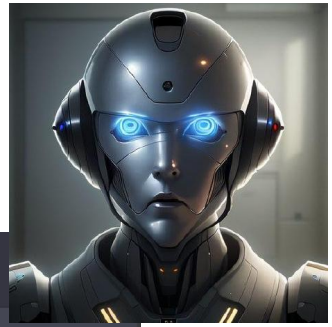
- Long-time Open Source contributor
  Ex-Gentoo dev, Sabayon, Ex-SUSE,
  Cloud foundry, openQA…
- Like Go, Hacking with big codebases,
  Perl Hacker
- I’ve created LocalAI, but also:
  kairos, edgevpn and luet

Github: @mudler

$agenda

- Understanding OpenAI and Generative AI
- Introduction to LocalAI (Use cases, Features and Capabilities, LocalAI vs OpenAI)
- Community and Open Source Development
- Architecture
- Model Families
- Challenges and Limitations
- Future Developments
- Q&A

$openai & chatgpt



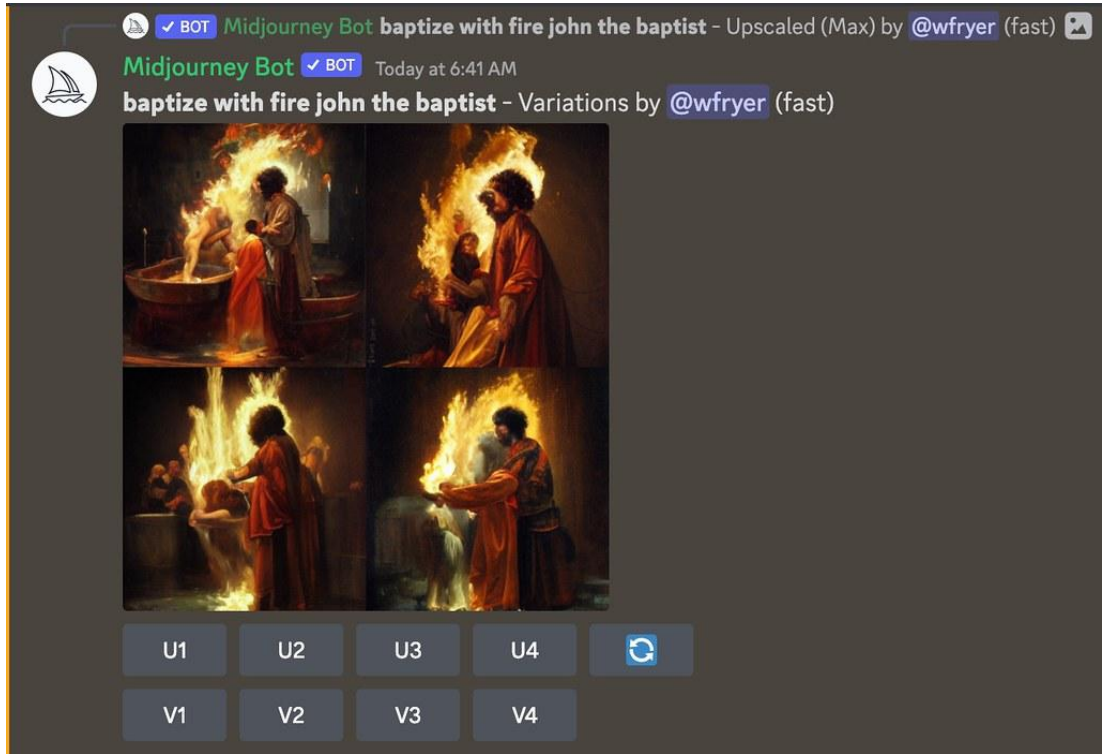
PO Che cos'è Wikipedia?

Wikipedia è un'enciclopedia online gratuita, multilingue e collaborativa che ha lo scopo di fornire informazioni affidabili su una vasta gamma di argomenti. Fondata nel 2001, Wikipedia è diventata una delle fonti di informazione più popolari in tutto il mondo e contiene milioni di articoli su una vasta gamma di argomenti, dalla storia alla tecnologia, dalla scienza alla cultura pop. La sua particolarità consiste nella possibilità per gli utenti di modificare e aggiornare gli articoli, il che significa che Wikipedia è costantemente in evoluzione e migliora costantemente la sua qualità e completezza. Tuttavia, poiché chiunque può modificare gli articoli, è importante verificare l'accuratezza delle informazioni fornite e utilizzare Wikipedia come punto di partenza per ulteriori ricerche piuttosto che come fonte definitiva.

Regenerate response

ChatGPT Jan 30 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

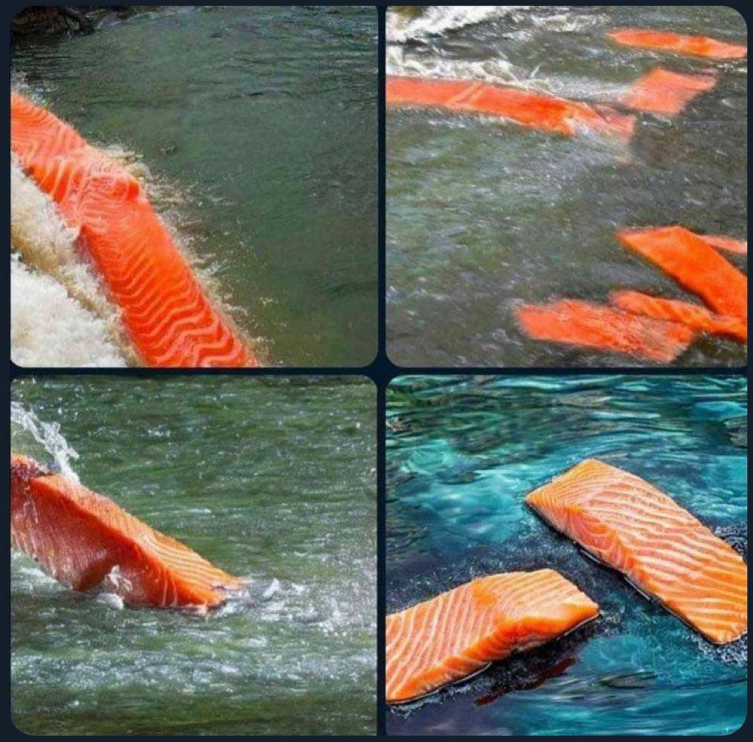
$generative AI



Voice cloning AI Dubbing ...

$generative AI

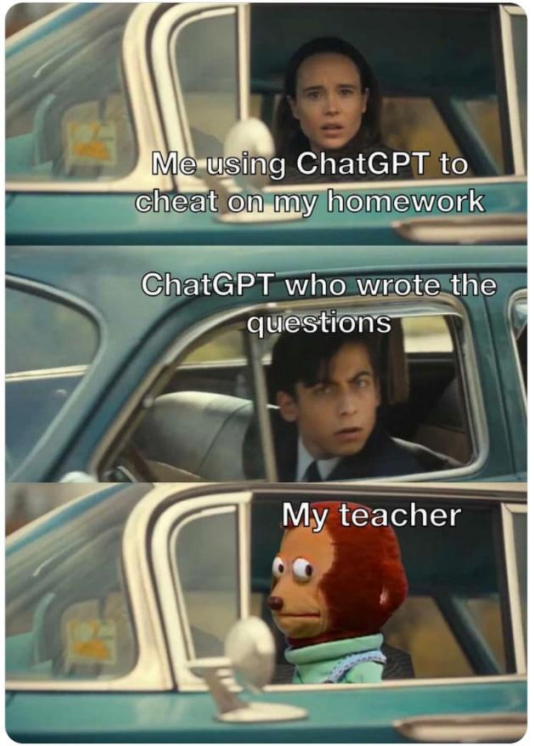
The AI prompt was “salmon in the river”. So majestic.



$impact

Sad & Useless Humor
@sadanduseless

The future of education.



2:29 AM · Jun 13, 2023

This.

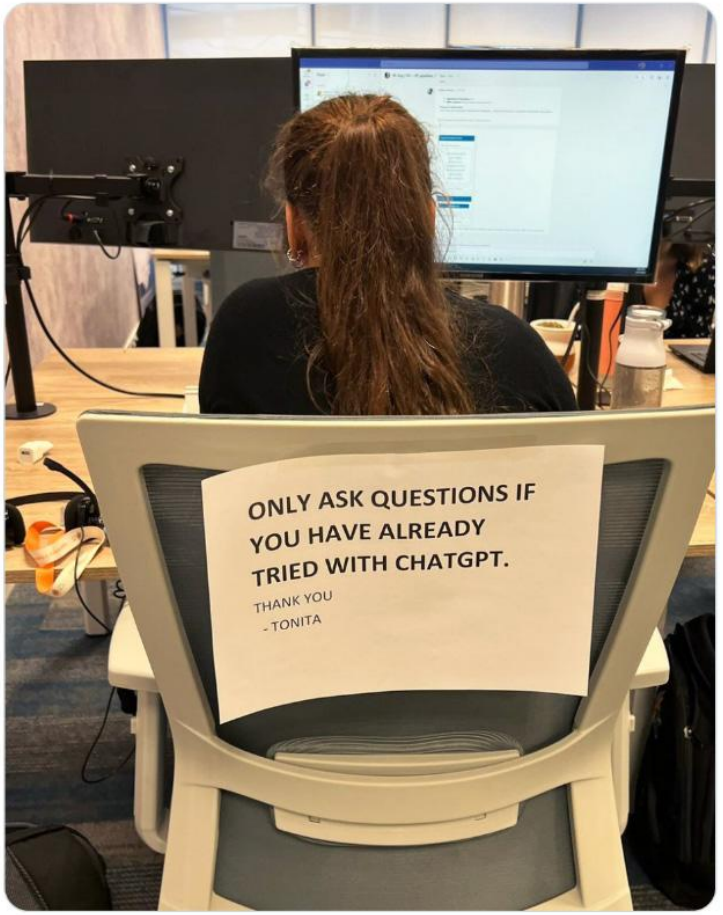
#ChatGPT will come for you too.

2:32 AM · Jun 13, 2023

2:25 AM · Jun 13, 2023

@sadanduseless

How to make your communication with coworkers less annoying.



2:56 AM · Jun 13, 2023

$open source LLM

Research use

Baize

Koala

Dalai

Lit-LLaMA

ColossalChat

Alpaca.cpp

LLaMA C++

Vicuna

Dolly

GPT4All

🦙🌲🤏 Alpaca-LoRA

Commercial use

BLOOMZ & mT0

Seldon

Flan-UL2

Cerebras-GPT

OpenChatKit

Pythia

nanoT5

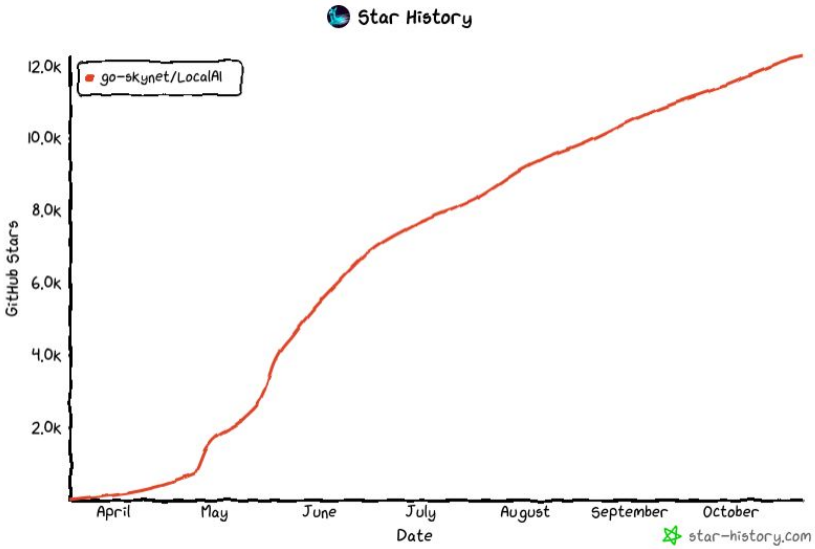
nanoGPT

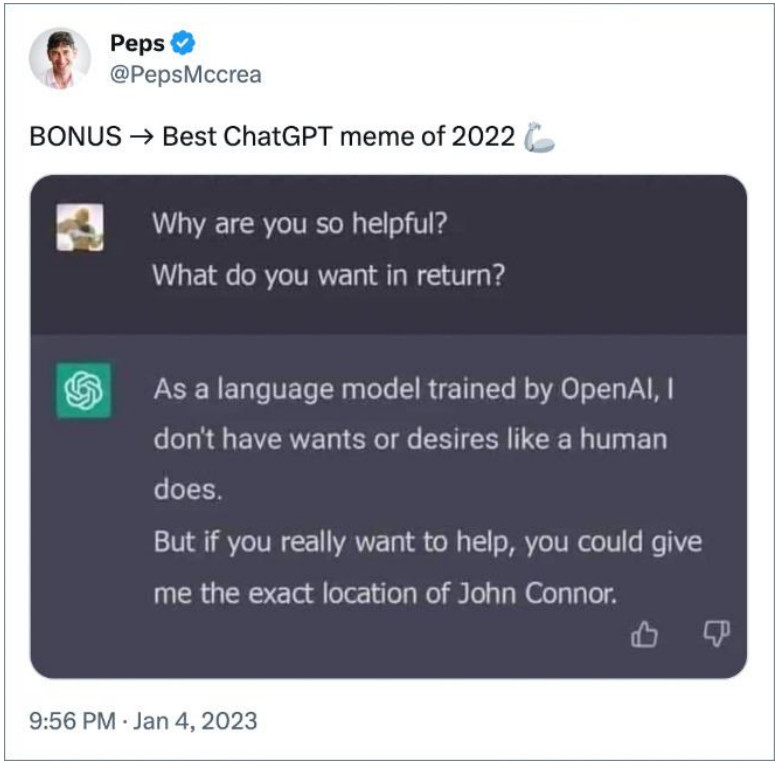
GeoV

By Sahar Mor

$LocalAI

How it started: a funny weekend project





$LocalAI vs OpenAI

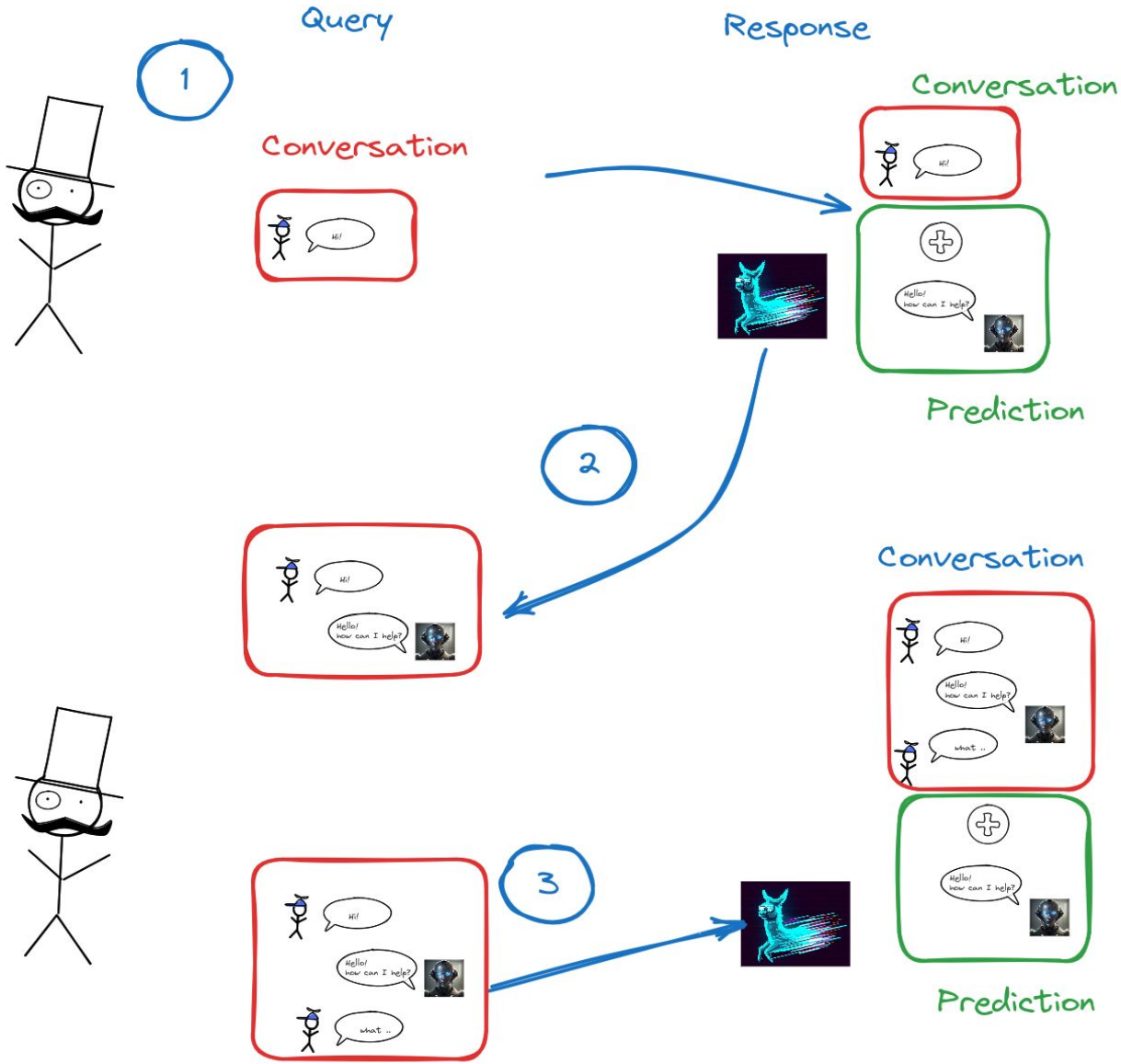
- Text generation (GPT)
- Text to Audio
- Audio to text
- Image generation
- Embeddings
- OpenAI functions
- Constrained grammars

Features	LocalAI	OpenAI
Functions	yes	yes
TTS	yes	no
Embeddings	yes	yes
Audio Transcriptions	yes	yes
Text generation	yes	yes
Image generation	yes	yes
Image edit	yes	no
Fine tuning	not yet	yes
Self-hostable	yes	no
Open source	yes	no

$usage

```
$> mkdir models
$> cp your-model.bin models/
$> docker run -p 8080:8080 -v $PWD/models:/models -ti --rm
quay.io/go-skynet/local-ai:latest --models-path /models

$> curl http://localhost:8080/v1/completions -H "Content-Type:
application/json" -d '{
     "model": "your-model.bin",
     "prompt": "A long time ago in a galaxy far, far away",
     "temperature": 0.7
   }'
$>
```

$API

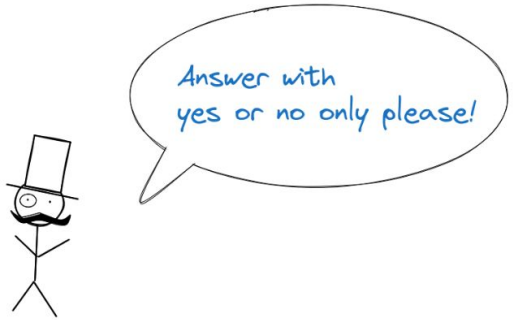

Query
Response
1
Conversation
Hi!
Conversation
Hi!
Hello! how can I help?
Prediction
2
Hi!
Hello! how can I help?
Conversation
Hi!
Hello! how can I help?
what ..
Hello! how can I help?
Prediction
3
Hi!
Hello! how can I help?
what ..

$Backus-Naur Form Grammars

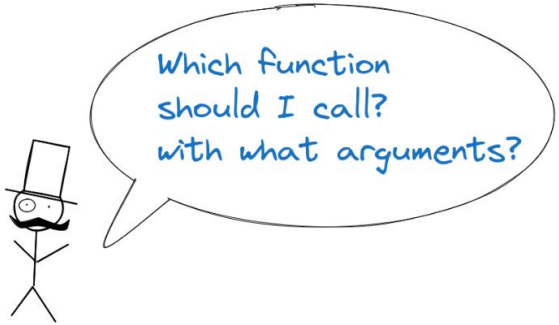
- Meta-syntax to define languages
- A set of derivation rules
- Used to describe syntax and languages (e.g. SQL)

```
<indirizzo postale> ::= <destinatario> <indirizzo> <localita>

<destinatario> ::= [<titolo>] [<nome>|<iniziale>] <cognome> <a capo>

<indirizzo> ::= <via> <numero civico> <a capo>

<localita> ::= [<CAP>] <comune> <provincia>
```

$Backus-Naur Form Grammars

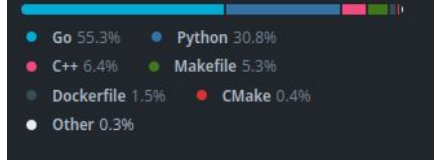


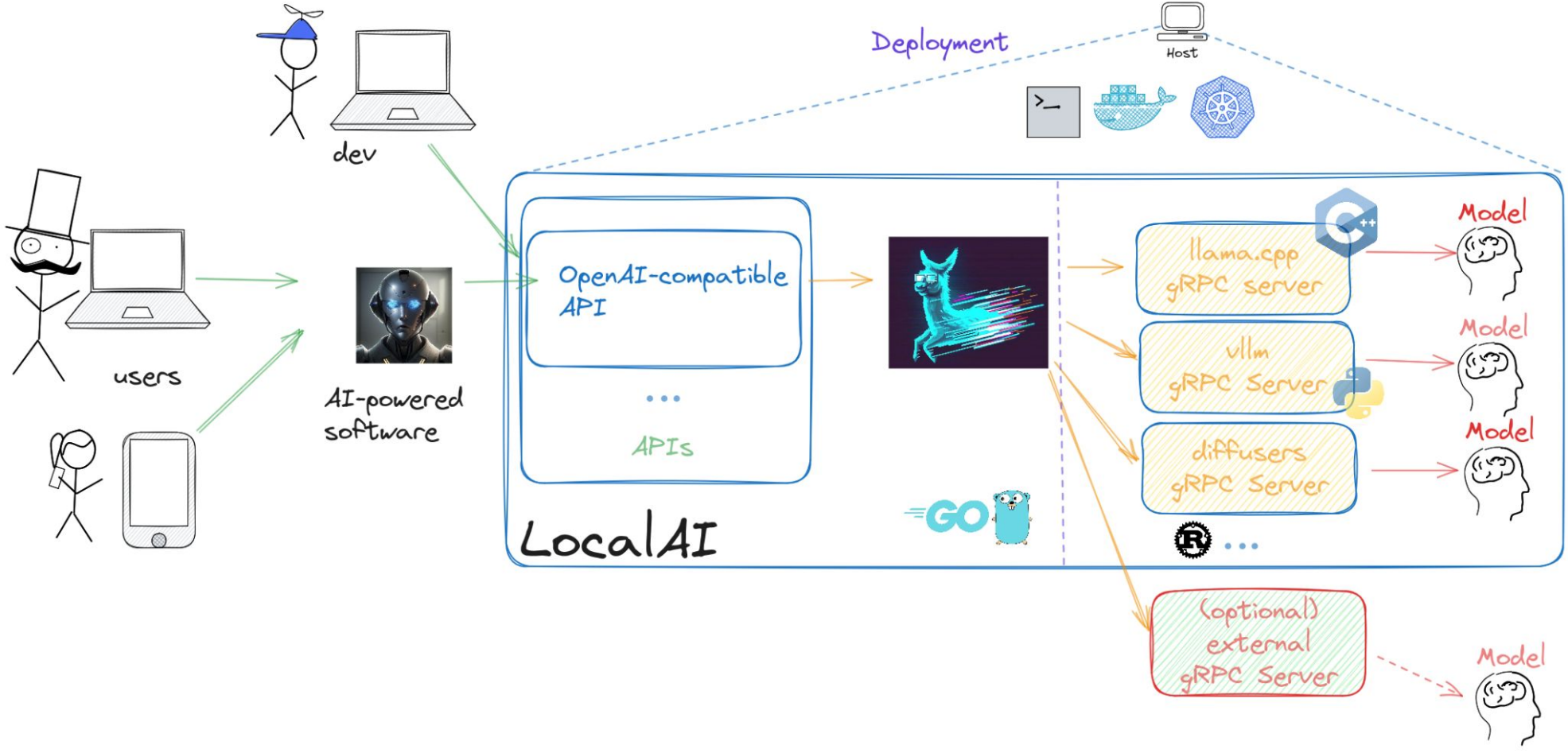

```
{
  "model": "gpt-4",
  "messages": [
      {"role": "user", "content": "Do you like apples?"}
  ],
  "grammar": "root ::= ("yes" | "no")"
}
```




```
{
  "model": "gpt-4",
  "messages": [
      {"role": "user", "content": "eat an apple!"}
  ],
  "functions": [
    "eat"
  ]
}
```

$Architecture





$use cases

Funny Pictures
@eatliver

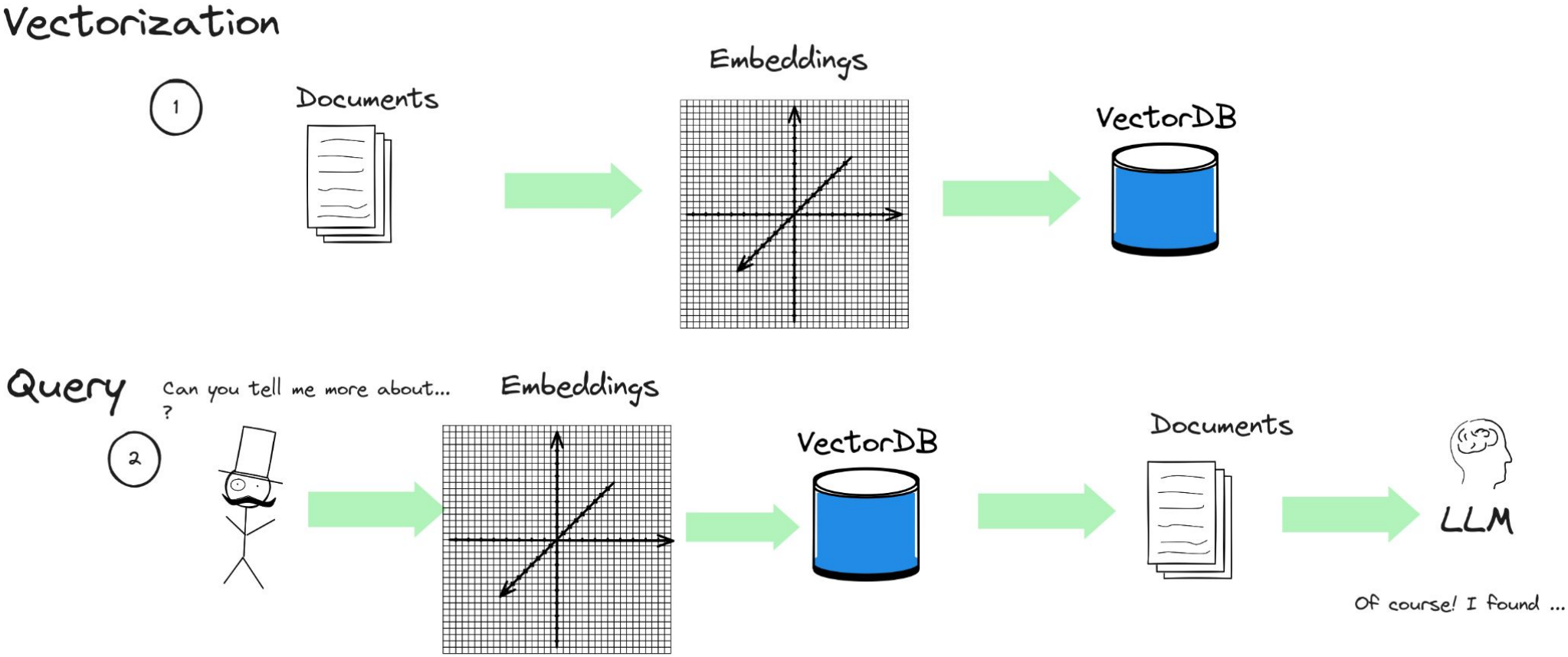
Yours sincerely, ChatGPT.

2:19 AM · Jun 13, 2023

$use cases

- Data conversion
- Virtual assistant
- Voice to text
- Text to voice
- Voice cloning
- Embeddings
- RAG systems

$use cases - RAG

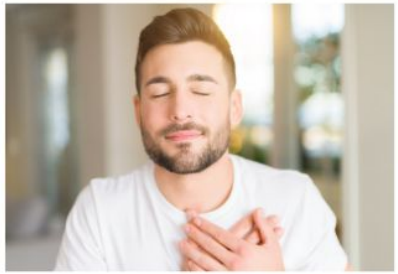


$local-llm community

https://localai.io/models/

How it started...

Thank you so much for this model, this is all we needed!



How it's going...

Wizard-Vicuna_Uncensored_Uncut_Extreme_Chaos_non-GMO_Organic-GGML.bin

Models 1,133 | gguf | new Full-text search | ↑↓ Sort: Trending

- TheBloke/zephyr-7B-alpha-GGUF — Updated 7 days ago • 397 • 97
- TheBloke/OpenHermes-2-Mistral-7B-GGUF — Updated 5 days ago • 42 • 34
- TheBloke/dolphin-2.1-mistral-7B-GGUF — Updated 10 days ago • 121 • 53
- TheBloke/Mistral-7B-Instruct-v0.1-GGUF — Text Generation • Updated 23 days ago • 3.92k • 201
- TheBloke/Llama-2-7b-Chat-GGUF — Text Generation • Updated 7 days ago • 3.75k • 96
- TheBloke/Mistral-7B-OpenOrca-GGUF — Text Generation • Updated 19 days ago • 1.25k • 144
- TheBloke/MistralLite-7B-GGUF — Updated 2 days ago • 7 • 13
- TheBloke/Mistral-7B-v0.1-GGUF — Text Generation • Updated 23 days ago • 2.01k • 118
- Undi95/Xwin-MLewd-13B-V0.2-GGUF — Updated 7 days ago • 10
- TheBloke/Xwin-LM-13B-v0.2-GGUF — Updated 7 days ago • 3 • 10
- IlyaGusev/saiga_mistral_7b_gguf — Conversational • Updated 12 days ago • 19
- TheBloke/Mistral-7B-Code-16K-qlora-GGUF — Updated 4 days ago • 14 • 8
- TheBloke/Mistral-7B-Phibrarian-32K-GGUF — Updated 4 days ago • 4 • 8
- TheBloke/llemma_7b-GGUF — Updated 4 days ago • 2 • 8
- TheBloke/CodeLlama-7B-GGUF — Text Generation • Updated 24 days ago • 264 • 41
- TheBloke/Phind-CodeLlama-34B-v2-GGUF — Updated 24 days ago • 181 • 80

$integrations

https://localai.io/integrations/

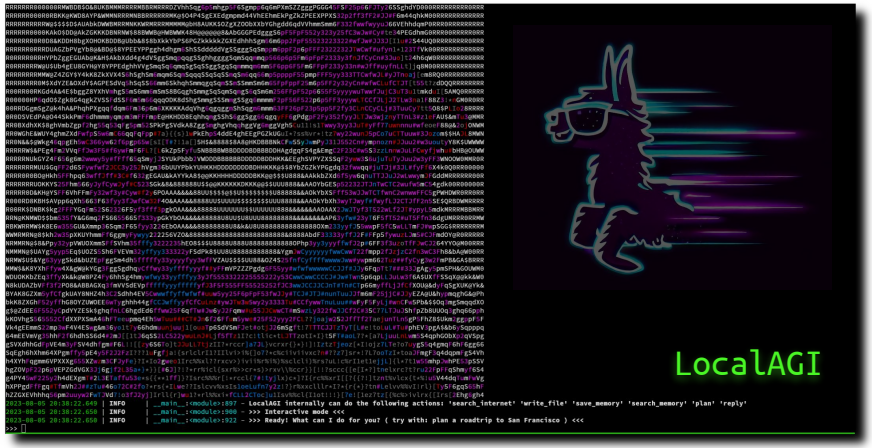
Integrations

- AutoGPT4all
- BionicGPT
- BMO Chatbo
- Flowise
- k8sgpt
- Kairos
- LinGoose
- LLMStack
- LocalAGI
- Mattermost-OpenOps
- Mods
- Spark

Many software already integrates natively with LocalAI:
the only requirement is that they are compatible with OpenAI and allow to configure an OpenAI remote host!

$integrations - LocalAGI

https://localai.io/integrations/



https://github.com/mudler/LocalAGI

https://github.com/mudler/LocalAGI/assets/2420543/23199ca3-7380-4efc-9fac-a6bc2b52bdb3

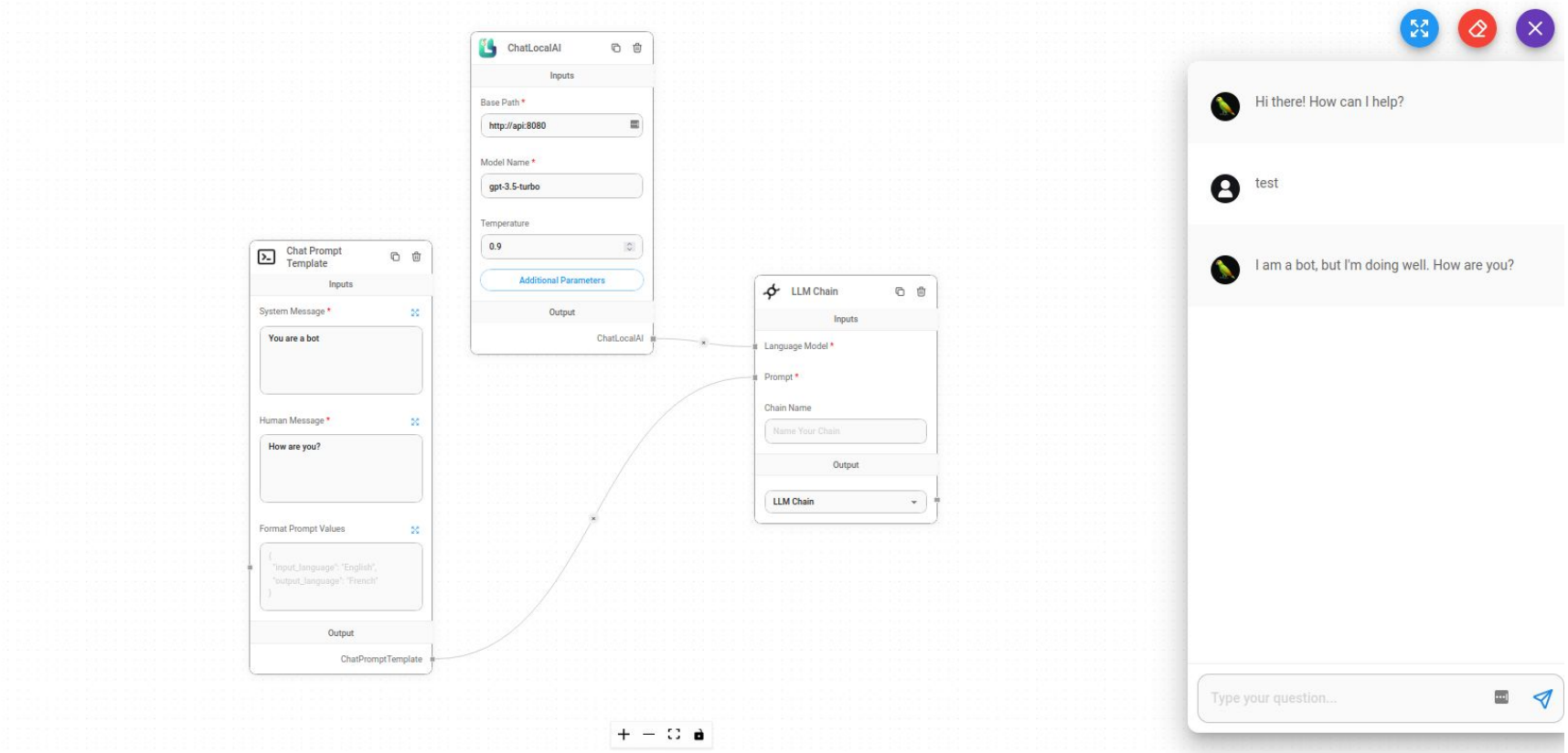
Highly customizable agent made for LocalAI:

- Embeddable library
- Uses functions
- Compatible with OpenAI APIs
- With voice, and avatar generation!

$integrations

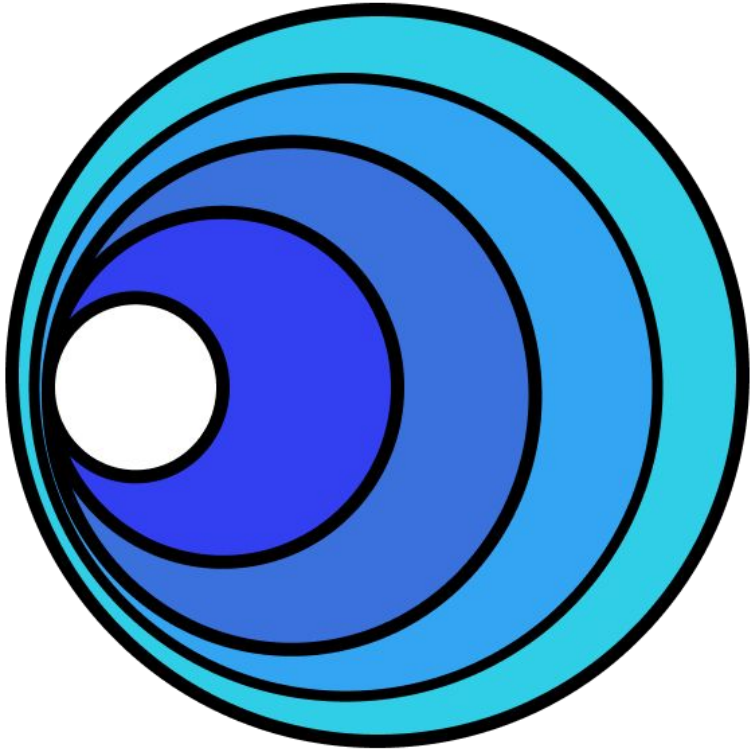
https://localai.io/integrations/

FlowiseAI



$integrations - k8sgpt

https://localai.io/integrations/



- SRE superpowers for everyone!
- Kubernetes analysis tool
- Integrates natively with LocalAI for in-cluster analysis

$future

🔥🔥 Hot topics / Roadmap 🔗

- [ ] Enable gallery management directly from the webui. https://github.com/go-skynet/LocalAI/issues/918
- [x] llama.cpp lora adapters: https://github.com/go-skynet/LocalAI/issues/919
- [ ] image variants and edits: https://github.com/go-skynet/LocalAI/issues/921
- [ ] esrgan for diffusers: https://github.com/go-skynet/LocalAI/issues/917
- [ ] ggml-stablediffusion: https://github.com/go-skynet/LocalAI/issues/916
- [ ] SAM: https://github.com/go-skynet/LocalAI/issues/915
- [x] diffusers lora adapters: https://github.com/go-skynet/LocalAI/issues/914
- [ ] resource management and control: https://github.com/go-skynet/LocalAI/issues/912
- [ ] ChatGLM: https://github.com/go-skynet/LocalAI/issues/754
- [ ] text-to-video : https://github.com/go-skynet/LocalAI/issues/933
- [ ] rustformers: https://github.com/go-skynet/LocalAI/issues/939
- [x] Vall-e: https://github.com/go-skynet/LocalAI/issues/985
- [ ] Speculative sampling: https://github.com/go-skynet/LocalAI/issues/1013
- [ ] Falcon/GPTNeoX on llama.cpp: https://github.com/go-skynet/LocalAI/issues/1009
- [ ] transformers/vllm: https://github.com/go-skynet/LocalAI/issues/1015
- [ ] TortoiseTTS: https://github.com/go-skynet/LocalAI/issues/1016
- [ ] Exllama2: https://github.com/go-skynet/LocalAI/issues/1053
- [ ] ctransformers: https://github.com/go-skynet/LocalAI/issues/1056
- [ ] GPTQ for LLama: https://github.com/go-skynet/LocalAI/issues/1055
- [ ] LLaVA and miniGPT-4: https://github.com/go-skynet/LocalAI/issues/1054
- [ ] Test generation inference: https://github.com/go-skynet/LocalAI/issues/1042
- [ ] docs - extending LocalAI with external backends: https://github.com/go-skynet/LocalAI/issues/1057

$Thank you!

Getting started:
https://localai.io/basics/getting_started/
Discord: https://discord.gg/uJAeKSAGDy
Github: https://github.com/mudler/LocalAI
Discussions:
https://github.com/mudler/LocalAI/discussions

Twitter: @mudler_it @LocalAI_API

$Q&A