
Ai
DIGEST

Основные новости

1. Выход GPT Store
2. Выход лучшей модели от Mistral - Mistral Medium и лучшей опенсорсной Mixtral
3. Выход Gemini от Google
4. Copilot разработчика от JetBrains
5. Обновленная MidJourney v6
6. Улучшенная генерация видео по фото от Runway
7. Новый сервис клонирования голоса OpenVoice

Нет доступа к OpenAI? Не беда!

...

Запускаем LLM на своем компьютере

Кто я?

– AI Product Director в Stealth Startup

– AI Solutions Product Lead в Wrike

– Head of Product в ABBYY

– Руководитель мобильных приложений в Сочи-2014

– Основатель мобильного стартапа

5 лет в AI, 9 лет в управлении продуктами

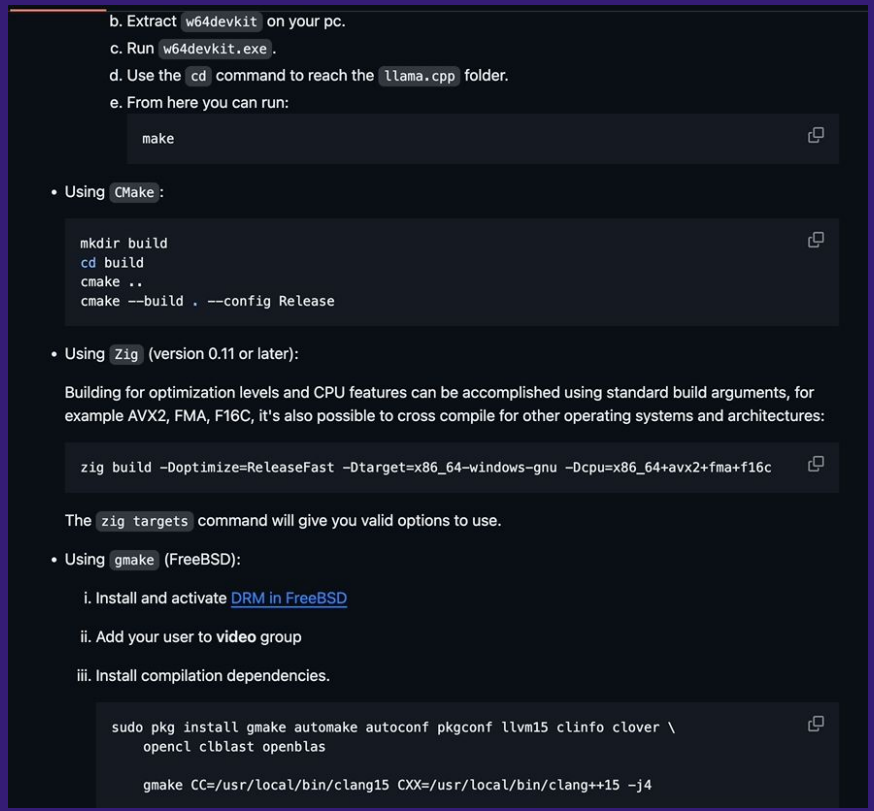
Ведущий канала AI в продукте - ai_product

Почему OpenAI не всегда помогает?

- Нет простого доступа к OpenAI ChatGPT из России
- Непросто купить доступ к премиуму российской картой
- На работе запрещен доступ
- Работаете с секретными и конфиденциальными данными

Что делать?

- Давайте запустим свою LLM, есть же опенсорс! :)
- Нет такой крутой видюхи :(
- Запустим сжатые модели :)
- Отлично, берем llama.cpp и ставим себе... Матерь божья... :(



b. Extract `w64devkit` on your pc.

c. Run `w64devkit.exe`.

d. Use the `cd` command to reach the `llama.cpp` folder.

e. From here you can run:

```
make
```

- Using `CMake`:

```
mkdir build
cd build
cmake ..
cmake --build . --config Release
```

- Using `Zig` (version 0.11 or later):

Building for optimization levels and CPU features can be accomplished using standard build arguments, for example AVX2, FMA, F16C, it's also possible to cross compile for other operating systems and architectures:

```
zig build -Doptimize=ReleaseFast -Dtarget=x86_64-windows-gnu -Dcpu=x86_64+avx2+fma+f16c
```

The `zig targets` command will give you valid options to use.

- Using `gmake` (FreeBSD):

i. Install and activate DRM in FreeBSD

ii. Add your user to **video** group

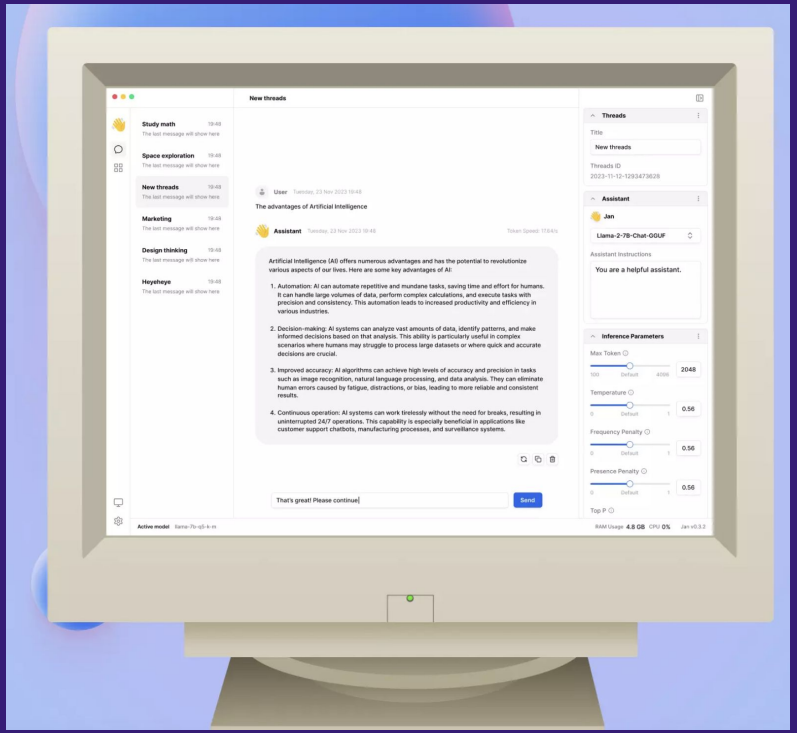
iii. Install compilation dependencies.

```
sudo pkg install gmake automake autoconf pkgconf llvm15 clinfo clover \
    opencl clblast openblas

gmake CC=/usr/local/bin/clang15 CXX=/usr/local/bin/clang++15 -j4
```

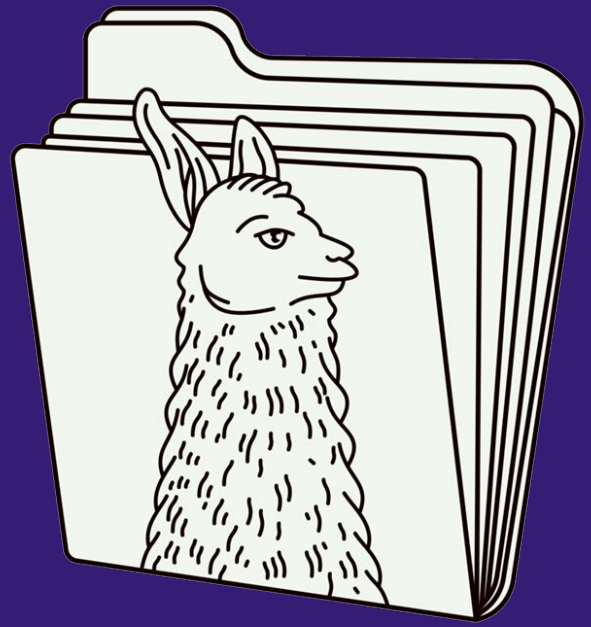
Ок, ищем готовые инсталляторы, лучше и под Mac, и под Win

- LlamaFile from Mozilla
- LM Studio
- Jan



LLaMaFile

1. Скачать нужный файл с GitHub
2. Дать права на выполнение (Mac, Linux) или добавить расширение *.exe (Windows)
3. Запустить файл (создать локальный сервер и откроет браузер со страничкой чата с выбранной моделью)
4. На текущий момент активно поддерживаются:



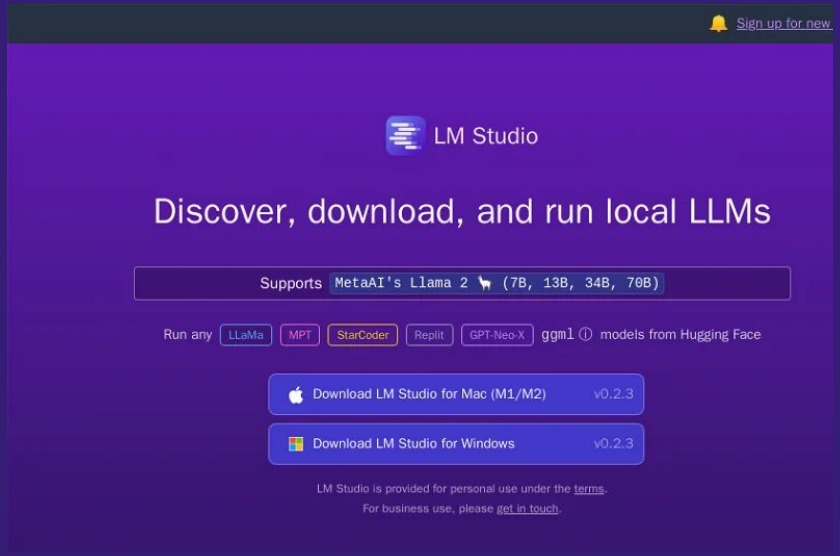
Model	Size	License	llamafile
LLaVA 1.5	3.97 GB	LLaMA 2	llava-v1.5-7b-q4.llamafile
Mistral-7B-Instruct	5.15 GB	Apache 2.0	mistral-7b-instruct-v0.2.Q5_K_M.llamafile
Mixtral-8x7B-Instruct	30.03 GB	Apache 2.0	mixtral-8x7b-instruct-v0.1.Q5_K_M.llamafile
WizardCoder-Python-13B	7.33 GB	LLaMA 2	wizardcoder-python-13b.llamafile

LM Studio

1. Скачать модуль

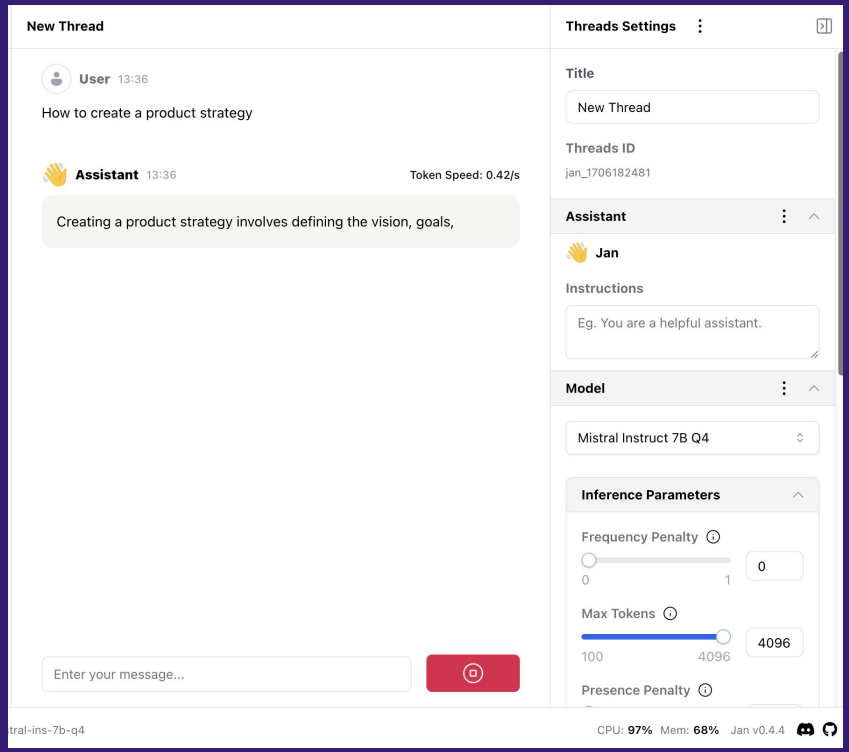
   (https://lmstudio.ai/)

2. Скачать нужные модели (много уже в самой оболочке, от LLama to Phi). А также можно на HuggingFace найти квантизованные модели или сделать нужную с помощью llama.cpp
3. Чаться (chat, instruct, code) в локальном окошке, не привлекая внимание (скажу сразу, без хорошей видеокарты будет медленно, на Mac с M1 - пойдет)



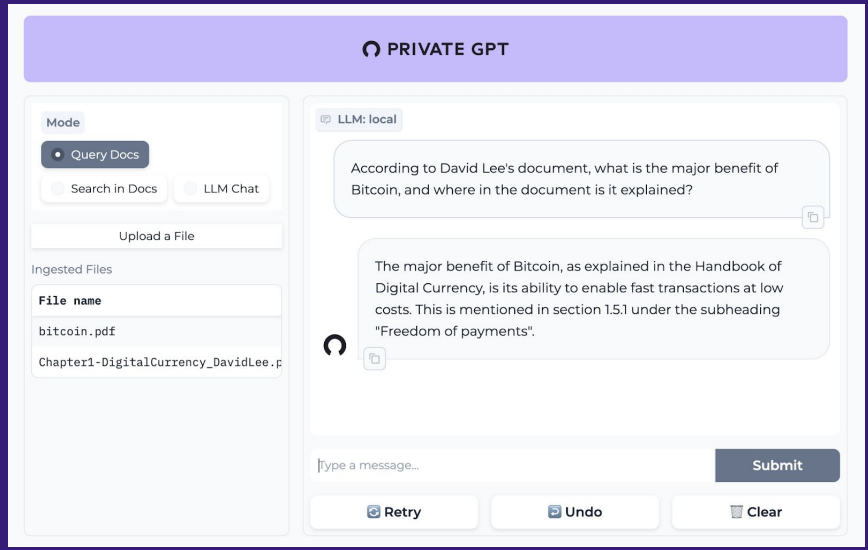
Jan

1. Open Source Desktop Client, скачать на Windows, Mac, Linux с jan.ai
2. Скачать нужные модели (если у вас слабый комп - mistral, phi, codewizard)
3. Чаться (chat, instruct, code) в локальном окошке, не привлекая внимание. Можно также подключить ключи к OpenAI через API, тогда на ваши данные не будут сохраняться и использоваться



Другие варианты, а также как развернуть свой чат с документами

- YakGPT - локальный доступ к ChatGPT через API
- Ollama WebUI - UI для локальных LLM
- GPT4ALL - Чат с загрузкой документа
- H2O GPT - QnA и чат, саммаризация загруженного документа
- PrivateGPT - Чат по документу, по документу
- Quivr - Ассистент по документам и приложениям
- BionicGPT - NoCode RAG с кучей интеграцией
- EmbedChain, AutoLLM и др.



Выводы

- Локальные модели развиваются и скоро можно будет запускать на своем компьютере или телефоне что-то уровня GPT-3.5, а на мощном компьютере - приближающуюся к GPT-4. Mac с M1 и выше - уже отлично, для Win желательно видеокарта с 6-8Gb VRAM, можно б/у
- Самая стабильная и простая в использовании оболочка - LM Studio
- Для RAG надо долго и мучительно выбирать из “терминал-реди” решений, но возможно скоро появятся и нормальные установщики с понятным и простым интерфейсом.
- Следите за новостями, каждую неделю расклад меняется.

Регистрируйтесь на онлайн-курс “AI в вашем продукте”

Научись применять AI в своем продукте: решать боли пользователей быстрее и "волшебнее" с помощью современных AI-инструментов, с реальным роадмапом от прототипа до продакшена.

Участникам вебинара подарок: промокод - aidea

Спасибо за внимание

Все материалы будут доступны тут:
https://t.me/aideameetup

Заходите на огонек:
TG Personal - @akimovpro
TG Channel - @ai_product

LinkedIn - https://www.linkedin.com/in/igorakimov1/