How-To **AI**

locally & offline

lora

llama

quantization

fine-tunning

Hugging Face

refinement

gguf / gptq

Vo svete AI je množstvo pojmov, zorientovať sa je náročné. My chceme len také domáce AI-kovanie.

But Why?

BECAUSE 🤷

Prečo by sme chceli rozbehať AI doma? Pretože možno máme doma na stole toto...

BECAUSE 🤷

...alebo toto...

But also (Mainly)

- hacker mindset
- working offline (no leak/store of queries to 3rd-party providers)
- SaaS LLMs change over time silently (censored, woke-ism)

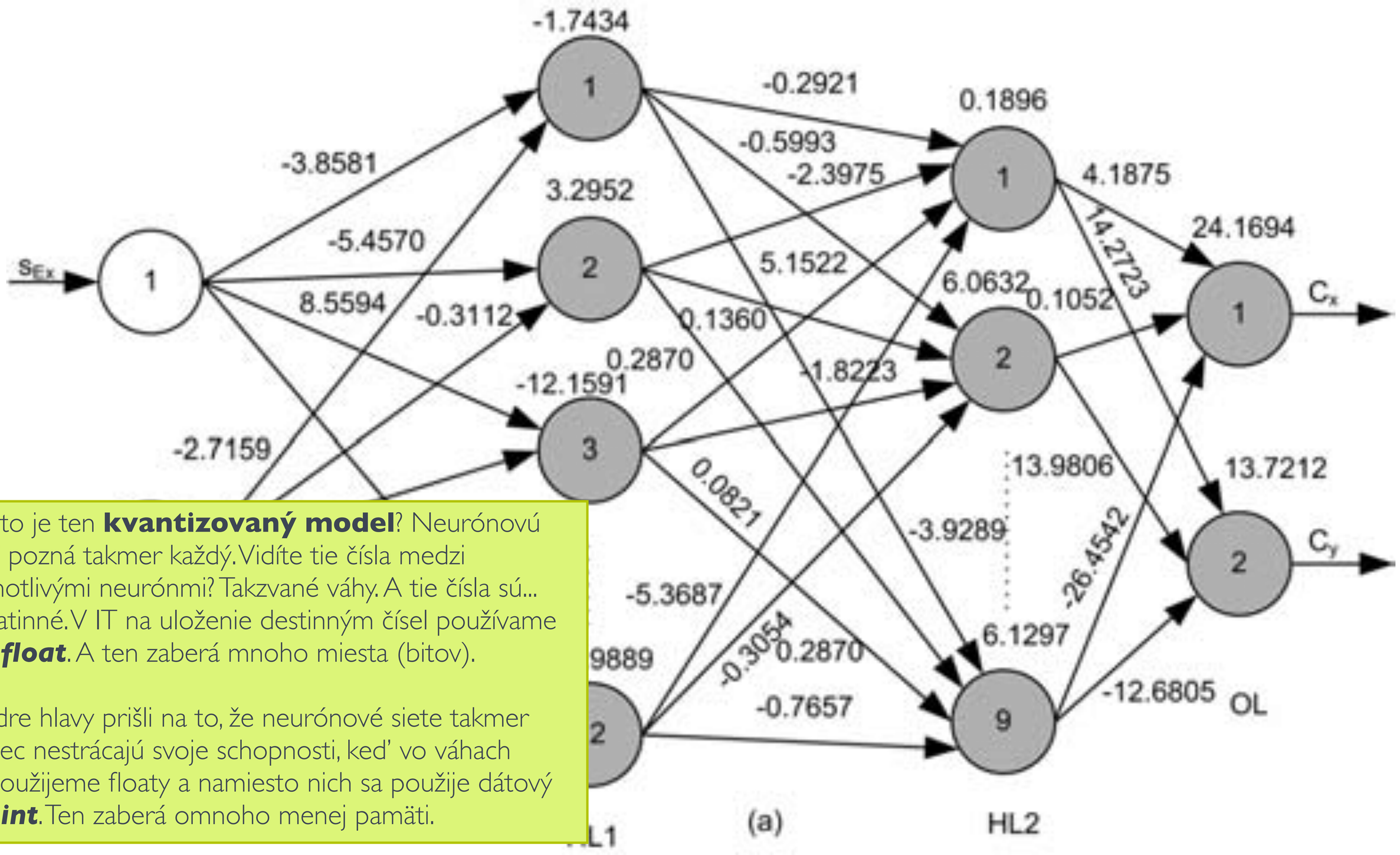
...ale hlavne pre tieto dôvody.

2 Groups of generative AIs

1. **LLMs** (ChatGPT)
2. **Image generation** (MidJourney)

HOW-TO START WITH **LLM**S

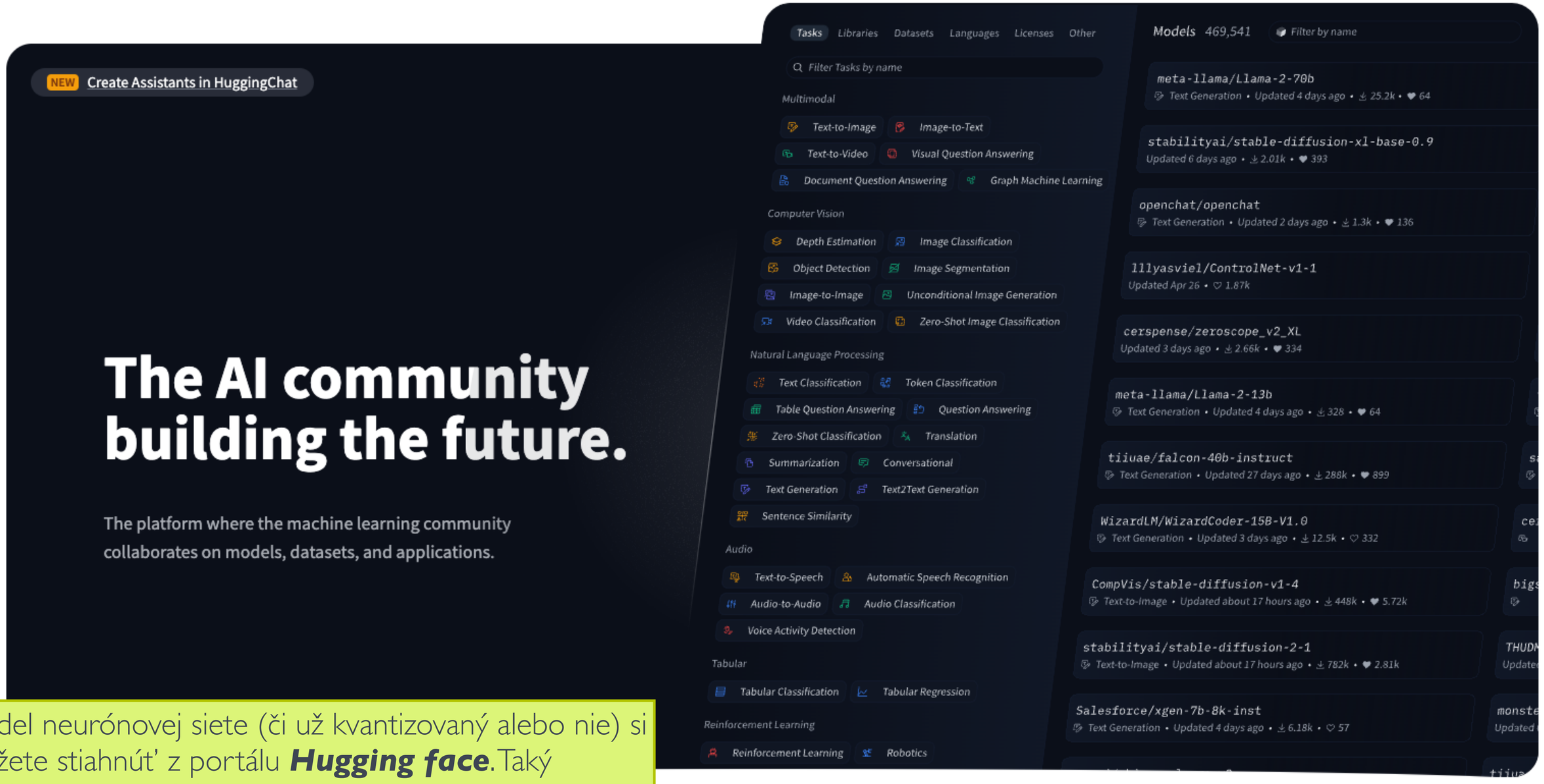
1. DOWNLOAD **QUANTIZED MODEL**
2. DOWNLOAD || COMPILE **RUNTIME**



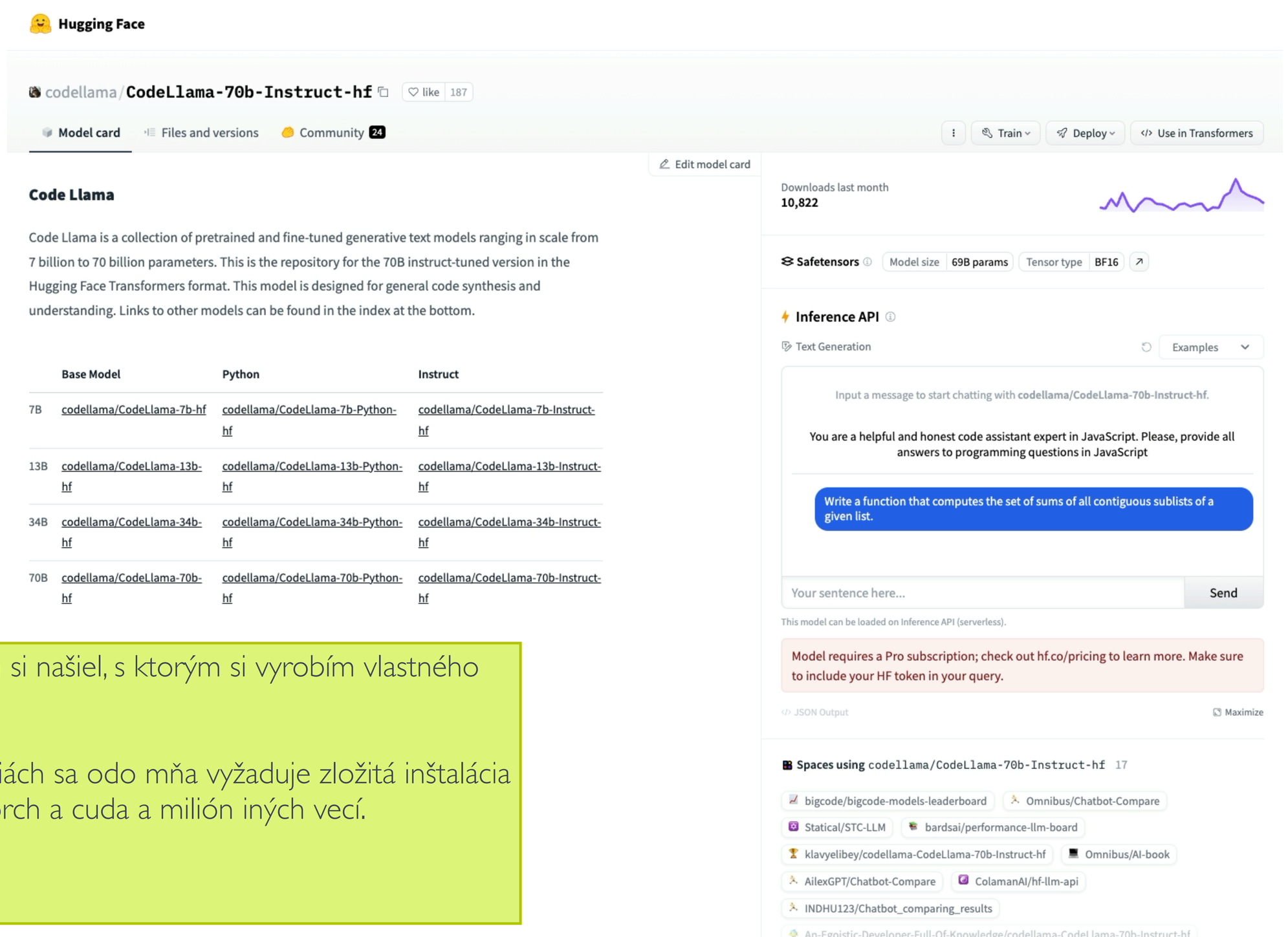
Čo to je ten **kvantizovaný model**? Neurónovú sieť pozná takmer každý. Vidíte tie čísla medzi jednotlivými neurónmi? Takzvané váhy. A tie čísla sú... desatinné. V IT na uloženie destinným čísel používame typ ***float***. A ten zaberá mnoho miesta (bitov).

Múdre hlavy prišli na to, že neurónové siete takmer vôbec nestrácajú svoje schopnosti, keď vo váhach nepoužijeme floaty a namiesto nich sa použije dátový typ ***int***. Ten zaberá omnoho menej pamäti.

Hugging Face



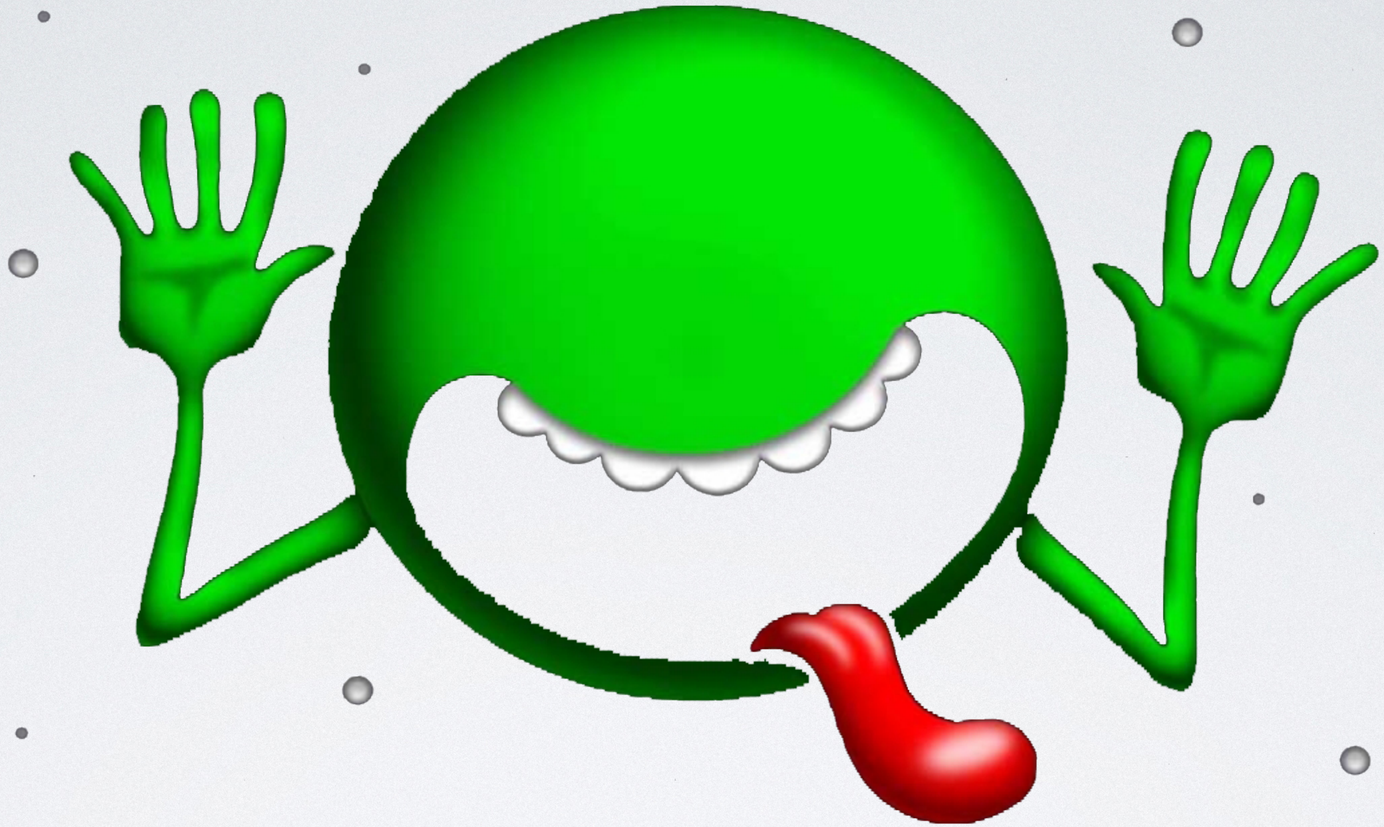
Model neurónovej siete (či už kvantizovaný alebo nie) si môžete stiahnúť z portálu ***Hugging face***. Taký GitHub pre neurónky.



Nojo, tak som si našiel, s ktorým si vyrobím vlastného ***copilot***a.

Ale v inštrukciách sa odo mňa vyžaduje zložitá inštalácia Pythonu, PyTorch a cuda a milión iných vecí.

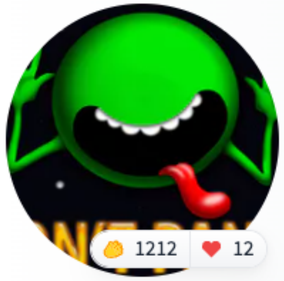
To nedám 😥



DON'T PANIC

Hugging Face

Na **HF** je tento... *týpek*. A on nerobí nič iné, iba celý deň kvantizuje existujúce modely a výsledok dáva späť na HF.



Tom Jobbins PRO

TheBloke

Follow

17769 followers · 16 following

TheBlokeAI TheBloke

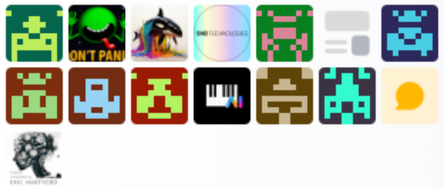
AI & ML interests

LLM: quantisation, fine tuning

Blog posts

Making LLMs lighter with AutoGPTQ and transf…
Aug 23, 2023

Organizations



Collections 1

Recent models: last 100 repos, sorted by creation date
The last 100 repos I have created. Sorted by creation date descending, so the most r…

- TheBloke/CapybaraHermes-2.5-Mistral-7B-GPTQ
  Updated Jan 31 • 3.68k • 31
- TheBloke/CapybaraHermes-2.5-Mistral-7B-AWQ
  Updated Jan 31 • 1.35k • 9
- TheBloke/CapybaraHermes-2.5-Mistral-7B-GGUF
  Updated Jan 31 • 162 • 50
- TheBloke/KafkaLM-70B-German-V0.1-AWQ

Models 3863

Sort: Recently updated

- TheBloke/CapybaraHermes-2.5-Mistral-7B-GPTQ
  Updated Jan 31 • 3.68k • 31
- TheBloke/CapybaraHermes-2.5-Mistral-7B-AWQ
  Updated Jan 31 • 1.35k • 9
- TheBloke/CapybaraHermes-2.5-Mistral-7B-GGUF
  Updated Jan 31 • 162 • 50
- TheBloke/KafkaLM-70B-German-V0.1-GPTQ
  Text Generation • Updated Jan 31 • 13 • 1
- TheBloke/KafkaLM-70B-German-V0.1-AWQ
  Text Generation • Updated Jan 31 • 24 • 2
- TheBloke/KafkaLM-70B-German-V0.1-GGUF
  Text Generation • Updated Jan 31 • 44 • 5
- TheBloke/CodeLlama-70B-Python-GPTQ
  Text Generation • Updated Jan 31 • 228 • 15
- TheBloke/CodeLlama-70B-Python-AWQ
  Text Generation • Updated Jan 31 • 84 • 5
- TheBloke/CodeLlama-70B-Python-GGUF
  Text Generation • Updated Jan 31 • 34 • 29
- TheBloke/CodeLlama-70B-Instruct-GPTQ
  Text Generation • Updated Jan 31 • 202 • 7

Expand 3863 models

TheBloke/Samantha-1.11-CodeLlama-34B-GPTQ
Text Generation • Updated 2 days ago • ⇩ 174 • ♡ 19

TheBloke/Samantha-1.11-
Updated 2 days ago • ⇩ 4 • ♡ 5

No dobre, ale ktorý si mám stiahnúť, a čo znamenajú tie prípony na konci?

TheBloke/CodeLlama-34B-GPTQ
Text Generation • Updated 2 days ago • ⇩ 561 • ♡ 15

TheBloke/Samantha-1.11-70B-GPTQ
Text Generation • Updated 2 days ago • ⇩ 54 • ♡ 3

GGML 🙅‍♂️ · old format, not supported by LLAMA.cpp

TheBloke/CodeLlama-34B-Instruct-GPTQ
Text Generation • Updated 2 days ago • ⇩ 7.47k • ♡ 36

TheBloke/CodeLlama-34B-Python-GPTQ
Text Generation • Updated 2 days ago • ⇩ 178 • ♡ 11

GPTQ 👍 · for GPU acceleration (multi-quantization)

TheBloke/CodeLlama-13B-Python-GPTQ

TheBloke/CodeLlama-7B-GPTQ

GGUF 👍 · for METAL acceleration (Mac)

TheBloke/CodeLlama-7B-Python-GPTQ
Text Generation • Updated 2 days ago • ⇩ 1.9k • ♡ 21

TheBloke/CodeLlama-34B-Instruct-GGUF
Text Generation • Updated 2 days ago • ⇩ 143 • ♡ 63

AWQ 😩 · mem. efficient GPU acceleration (nVidia only)

Text Generation • Updated 2 days ago • ⇩ 38 • ♡ 29

Text Generation • Updated 2 days ago • ⇩ 59 • ♡ 43

TheBloke/CodeLlama-13B-Python-GGUF
Text Generation • Updated 2 days ago • ⇩ 52 • ♡ 18

TheBloke/CodeLlama-7B-Instruct-GPTQ
Text Generation • Updated 2 days ago • ⇩ 4.71k • ♡ 38

TheBloke/CodeLlama-13B-Instruct-GGUF

TheBloke/CodeLlama-13B-GGUF

OK, keď nájdeš vhodný model pre svoj HW, na jeho detail stránke nájdeš pekný popis. Zoscrolluj dole na sekciu ***Provided files***...

Hugging Face

TheBloke/**CodeLlama-13B-Instruct-GGUF** like 45

Text Generation | Transformers | code | llama | llama-2 | text-generation-inference | arxiv:2308.12950 | License: llama2

Model card | Files | Community 2 | Train | Deploy | Use in Transformers

Edit model card

Chat & support: TheBloke's Discord server

Want to contribute? TheBloke's Patreon page

TheBloke's LLM work is generously supported by a grant from andreessen horowitz (a16z)

CodeLlama 13B Instruct - GGUF

- Model creator: Meta
- Original model: CodeLlama 13B Instruct

Downloads last month
212

Hosted inference API

Text Generation

Inference API has been turned off for this model.

Space using TheBloke/CodeLlama-13B-Instru... 1

mikeee/codellama-13b-instruct-gguf

Provided files

Typek model kvantizuje na rôzny rozsah dátového typu ***int***. Napr. iba na 2 alebo 3 bity. To ovplyvňuje kvalitu modelu, ale aj... koľko pamäti bude tvoj HW potrebovať.

Name	Quant method	Bits	Size	Max RAM required	Use case
codellama-13b-instruct.Q2_K.gguf	Q2_K	2	5.43 GB	7.93 GB	smallest, significant quality loss - not recommended for most purposes
codellama-13b-instruct.Q3_K_S.gguf	Q3_K_S	3	5.66 GB	8.16 GB	very small, high quality loss
codellama-13b-instruct.Q3_K_M.gguf	Q3_K_M	3	6.34 GB	8.84 GB	very small, high quality loss
codellama-13b-instruct.Q3_K_L.gguf	Q3_K_L	3	6.93 GB	9.43 GB	small, substantial quality loss
codellama-13b-	Q4_0	4	7.37	9.87 GB	legacy; small,

codellama-13b-instruct.Q4_K_M.gguf	Q4_K_M	4	7.87 GB	10.37 GB	medium, balanced quality - recommended
codellama-13b-instruct.Q5_0.gguf	Q5_0	5	8.97 GB	11.47 GB	legacy; medium, balanced quality - prefer using Q4_K_M
codellama-13b-instruct.Q5_K_S.gguf	Q5_K_S	5	8.97 GB	11.47 GB	large, low quality loss - recommended
codellama-13b-instruct.Q5_K_M.gguf	Q5_K_M	5	9.23 GB	11.73 GB	large, very low quality loss - recommended
codellama-13b-instruct.Q6_K.gguf	Q6_K	6	10.68 GB	13.18 GB	very large, extremely low quality loss
codellama-13b-instruct.Q8_0.gguf	Q8_0	8	13.83 GB	16.33 GB	very large, extremely low quality loss - not

viz. iné kvantizácie...

Pozor, nemôžeš použiť všetkú pamäť svojho počítača!

MEMORY BUDGET GAME

24GB

Nejakú pamäť si ukúsne operačný systém...

MEMORY BUDGET GAME

...a Chrome (ktorý máš isto pustený celý deň)...

MEMORY BUDGET GAME

...a ďalšie aplikácie, ktoré používaš. Zisti si preto najskôr, koľko pamäti môžeš pre model vyhradiť.

MEMORY BUDGET GAME

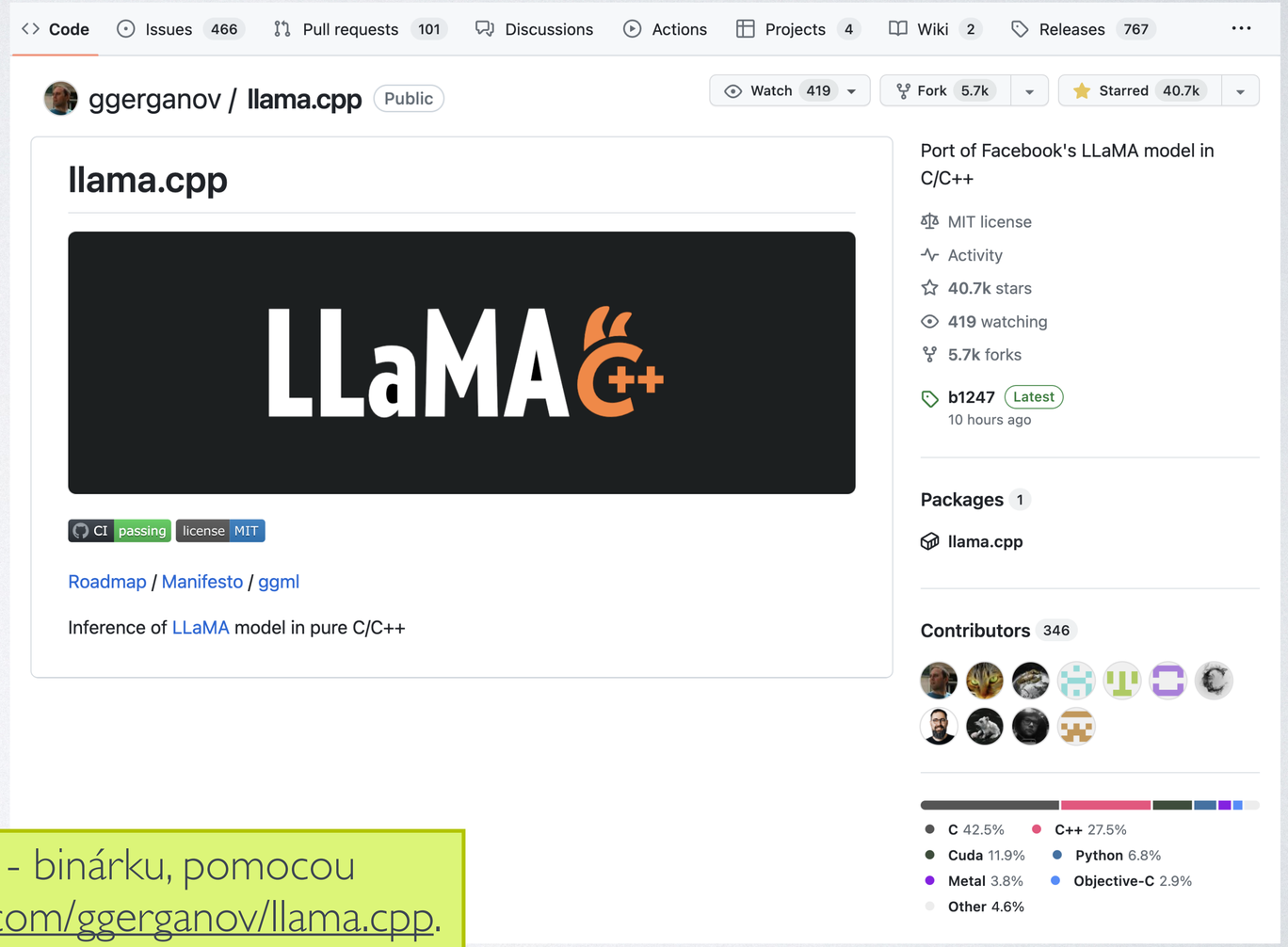
CodeLlama-**34B**-Instruct-GGUF

Modely s veľkým počtom váh vyžadujú naozaj veľké množstvo pamäti.

instruct.Q4_K_M.gguf			GB		balanced quality - recommended
codellama-34b-instruct.Q5_0.gguf	Q5_0	5	23.24 GB	25.74 GB	legacy; medium, balanced quality - prefer using Q4_K_M
codellama-34b-instruct.Q5_K_S.gguf	Q5_K_S	5	23.24 GB	25.74 GB	large, low quality loss - recommended
codellama-34b-instruct.Q5_K_M.gguf	Q5_K_M	5	23.84 GB	26.34 GB	large, very low quality loss - recommended
codellama-34b-instruct.Q6_K.gguf	Q6_K	6	27.68 GB	30.18 GB	very large, extremely low quality loss
codellama-34b-instruct.Q8_0.gguf	Q8_0	8	35.86 GB	38.36 GB	very large, extremely low quality loss - not recommended

HOW-TO START WITH **LLM**S

1. DOWNLOAD **QUANTIZED MODEL**
2. DOWNLOAD || COMPILE **RUNTIME**



Okrem dátového súboru modelu, budeš potrebovať runtime - binárku, pomocou ktorej spustíš model na svojom HW. Na začiatok skús github.com/ggerganov/llama.cpp.

ggerganov / llama.cpp

b2548
e82f9e2
Compare

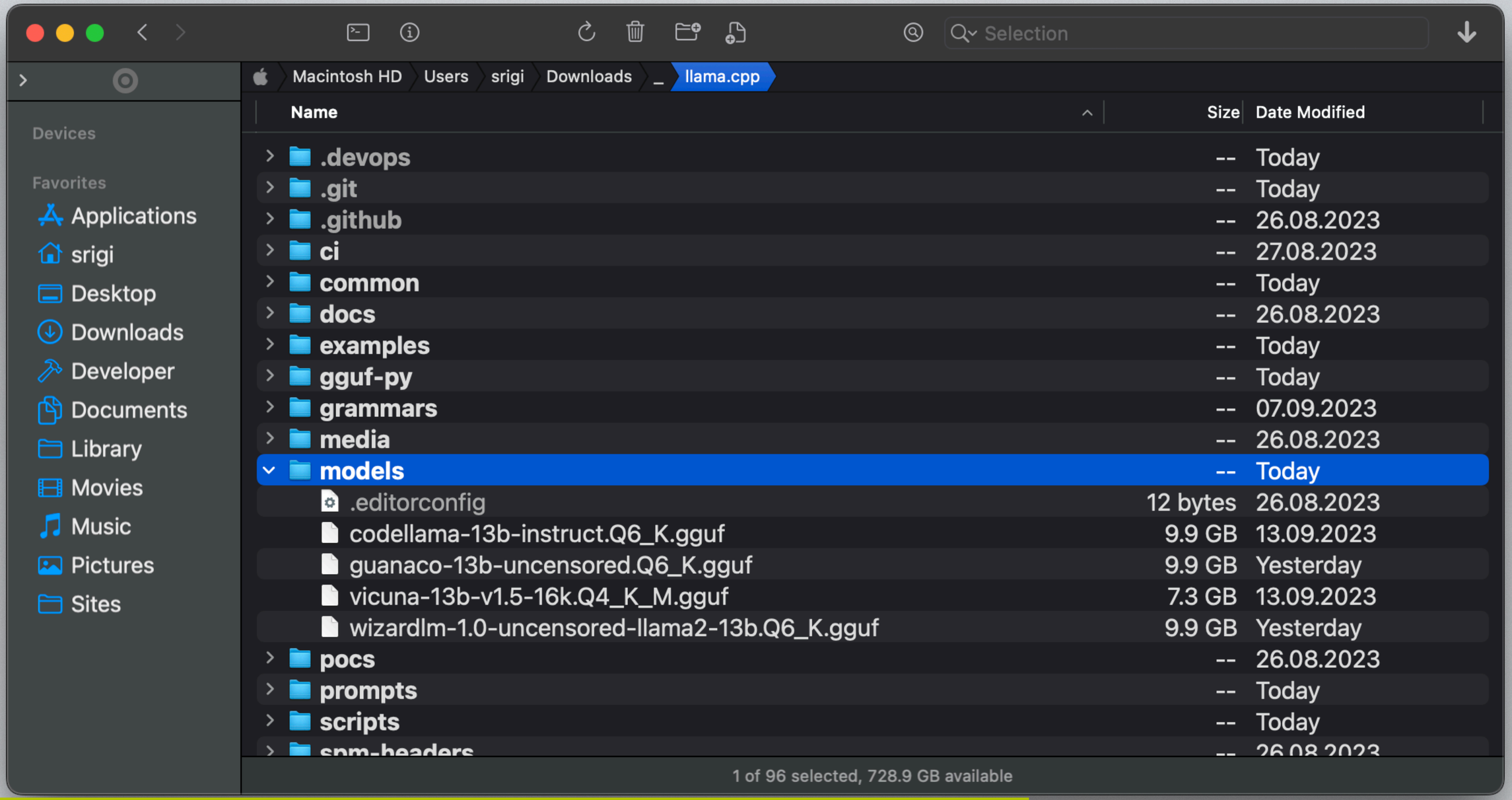
b2548 Latest

```
[SYCL] Fix batched impl for NVidia GPU (#6164)

* Fix batched impl

* Maintain previous behaviour for igpu

* retrigger CI
```

Assets 18

cudart-llama-bin-win-cu11.7.1-x64.zip	293 MB	1 hour ago
cudart-llama-bin-win-cu12.2.0-x64.zip	413 MB	1 hour ago
llama-b1-bin-macos-arm64.zip	37.1 MB	1 hour ago
llama-b1-bin-macos-x64.zip	41.8 MB	1 hour ago
llama-b2548-bin-win-arm64-x64.zip	4.32 MB	1 hour ago
llama-b2548-bin-win-avx-x64.zip	4.84 MB	1 hour ago
llama-b2548-bin-win-avx2-x64.zip	4.81 MB	1 hour ago
llama-b2548-bin-win-avx512-x64.zip	4.82 MB	1 hour ago
llama-b2548-bin-win-clblast-x64.zip	6.01 MB	1 hour ago
llama-b2548-bin-win-cuda-cu11.7.1-x64.zip	30.9 MB	1 hour ago
		4 hours ago
		4 hours ago

Netráp sa s nejakým kompilovaním. Na strákne ***Releases*** nájdeš predkompilované binárky pre takmer každý HW a s podporou akcelerácie.



Llama.cpp má pre modely predpripravenú zložku. Môžeš si tam hodiť aj viac modelov.

- GPTQ models for GPU inference, with multiple quantisation parameter options.
- 2, 3, 4, 5, 6 and 8-bit GGUF models for CPU+GPU inference
- KoboldAI's original unquantised fp16 model in pytorch format, for GPU inference and for further conversions

Prompt template: Alpaca

```
Below is an instruction that describes a task. Write a response that

### Instruction:
{prompt}

### Response:
```

Licensing

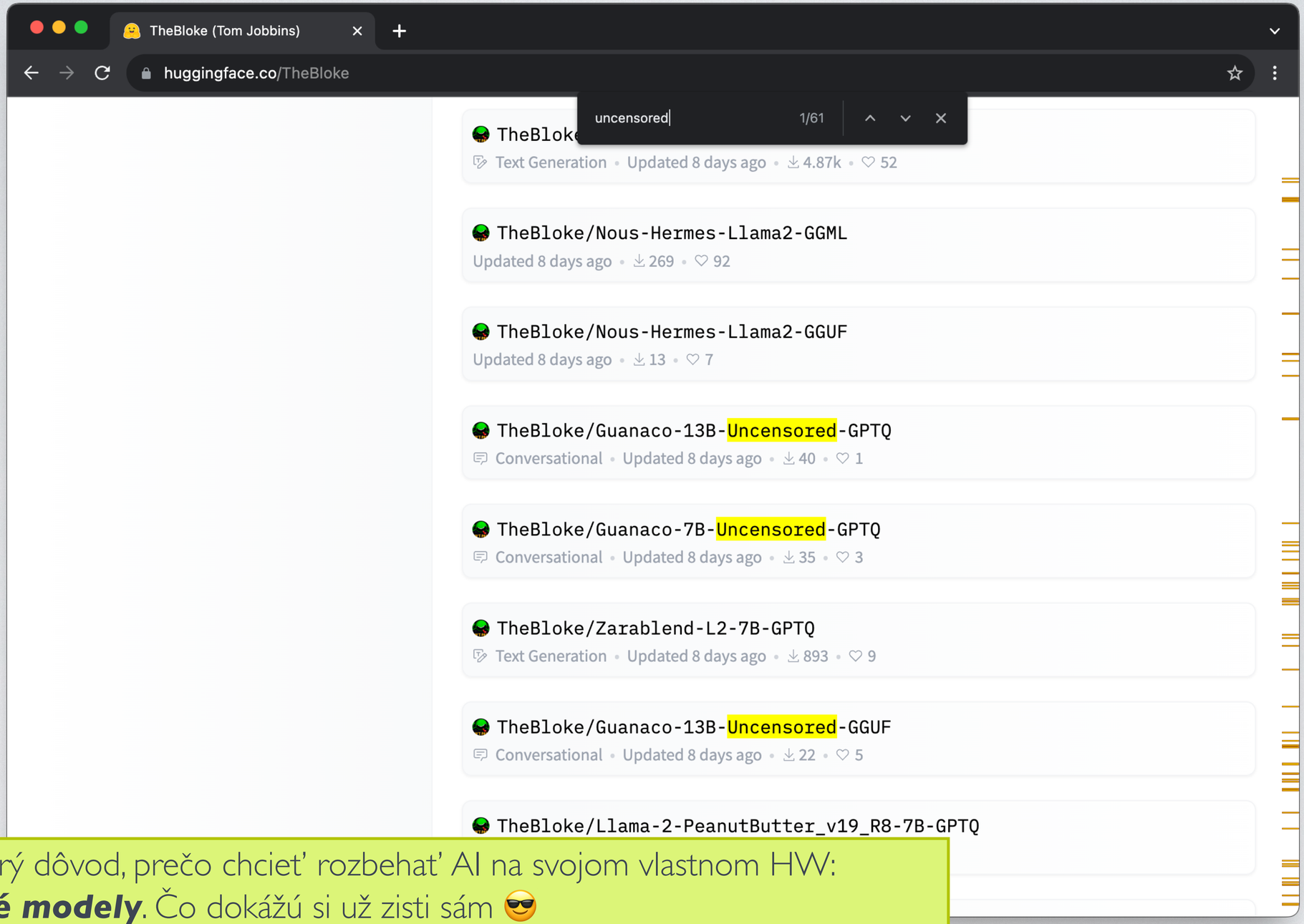
The creator of the source model has listed its license as `cc-by-nc-4.0`, and this quantization has therefore

As this model is based on L
and the license files for that are additionally included. It should therefore be

Na detail stránke modelu od *týpka* si ešte nájdi sekciu ***Prompt template***. Budeš to potrebovať hneď v ďalšom kroku.
Ukazuje ako model očakáva, že mu zadáš *prompt*.

```bash
#!/usr/bin/env bash

LLAMA_CPP_PATH=$HOME/Downloads/_/llama.cpp
MODEL="$LLAMA_CPP_PATH/models/llama2-13b-estopia.Q5_K_M.gguf"
MODEL_LAYERS=41

$LLAMA_CPP_PATH/main \
    --color \
    --escape \
    --log-disable \
    --model $MODEL \
    --n-gpu-layers $MODEL_LAYERS \
    --threads 8 \
    --ctx_size 6144 \
    --n-predict -1 \
    --interactive-first \
    --in-prefix "\n### Instruction: " \
    --in-suffix "\n### Response: " \
    --multiline-input \
    --prompt "Below is an instruction that describes a task. Write a response that appropriately completes the request.\n"
```

Vyrob si bash script (na Windowse nazývaný tiež *baťák*), ktorým spustíš llama.cpp. Na obrázku vidíš odladený script pre model *llama2*. Najdôležitejší je argument **n-gpu-layers**, ktorým určuješ koľko vrstiev neurónov pošleš na akceleráciu na GPU. Počet layerov modelu zistiš tak, že odblokuješ logovanie (viz. argument **log-disable**).
Všimni si, ako sme nakonfigurovali promptovanie. Je to celkom mágia, hore vidíš setup pre *prompt template* z predošlého screenu.
Pozor, je zapnutý **multiline-input** a preto musíš svoj prompt ukončiť znakom ****

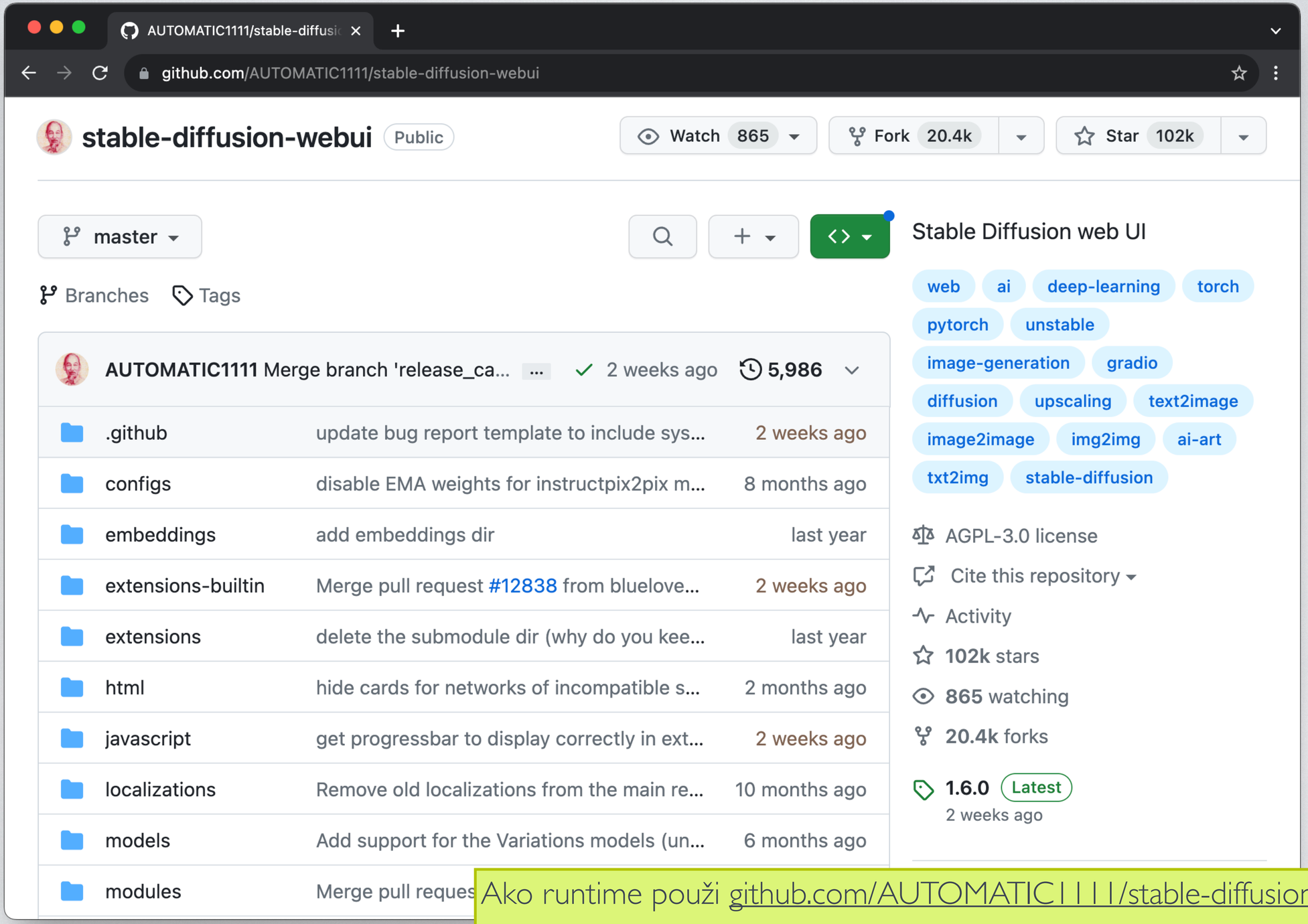


Je ešte jeden dobrý dôvod, prečo chcieť rozbehať AI na svojom vlastnom HW: ***necenzúrované modely***. Čo dokážú si už zisti sám 😎

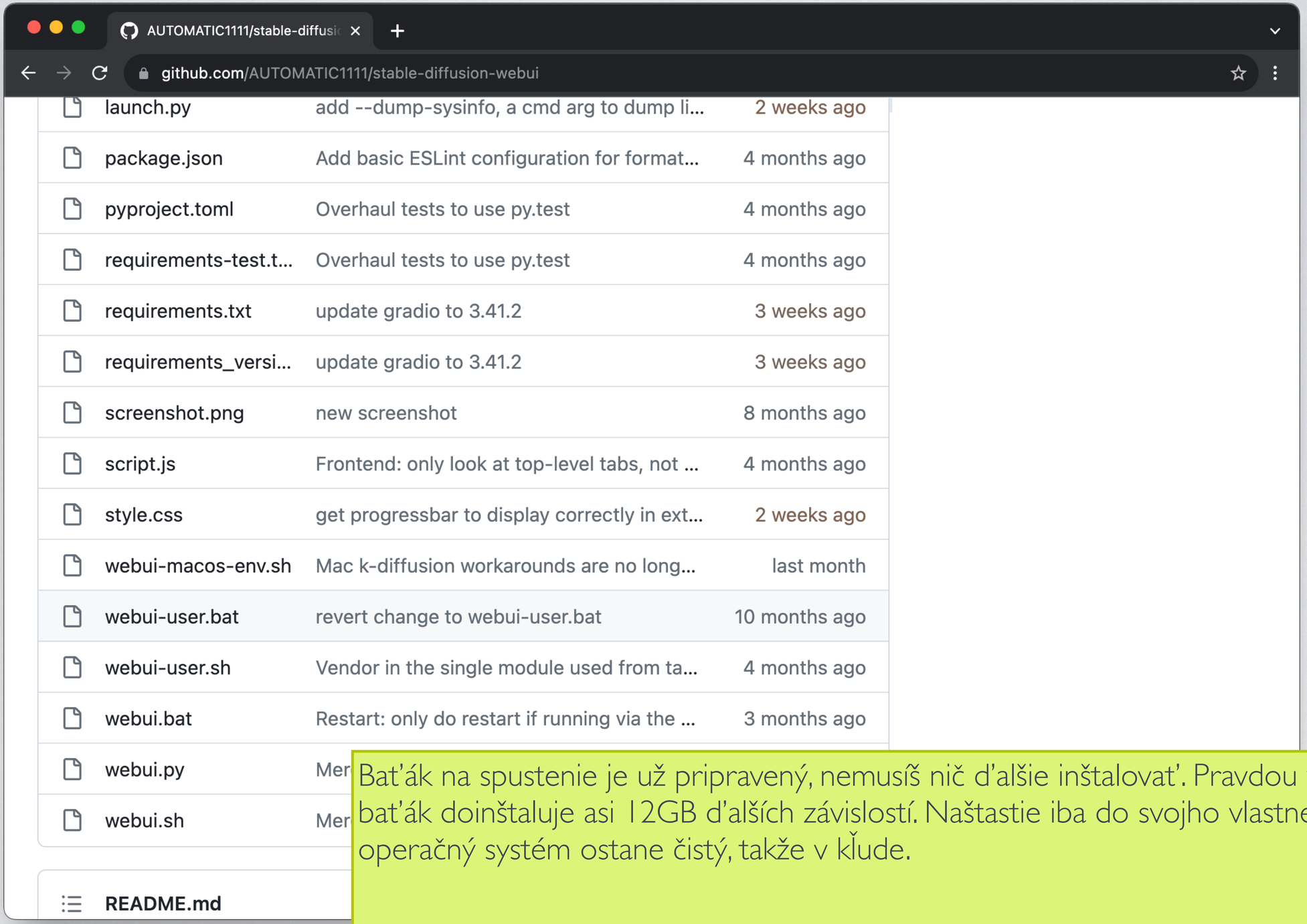
How-To start with **Image generation**

Generovanie obrázkov už vyžaduje dosť silný HW. S nejakým Macbook AIR nepochodíš.

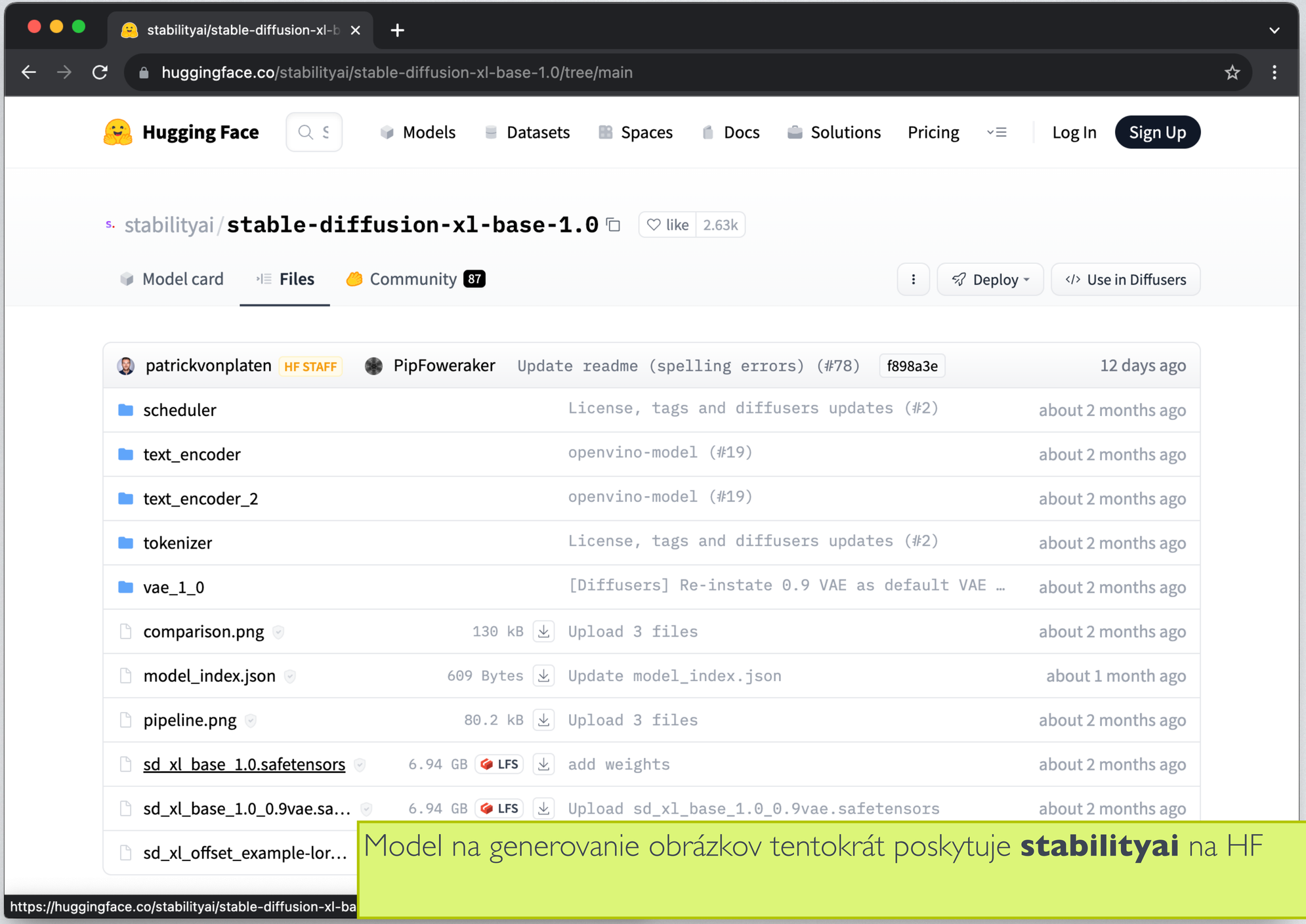
Ideálne je mať k dispozícii takúto hračku



Ako runtime použi github.com/AUTOMATIC1111/stable-diffusion-webui



Bat'ák na spustenie je už pripravený, nemusíš nič d'alšie inštalovať. Pravdou však je, že bat'ák doinštaluje asi 12GB d'alších závislostí. Naštastie iba do svojho vlastného adresára, operačný systém ostane čistý, takže v kľude.


stabilityai/stable-diffusion-xl-b
huggingface.co/stabilityai/stable-diffusion-xl-base-1.0/tree/main
Hugging Face
Models
Datasets
Spaces
Docs
Solutions
Pricing
Log In
Sign Up
stabilityai / stable-diffusion-xl-base-1.0
like 2.63k
Model card
Files
Community 87
Deploy
Use in Diffusers
patrickvonplaten HF STAFF PipFoweraker Update readme (spelling errors) (#78) f898a3e 12 days ago
scheduler License, tags and diffusers updates (#2) about 2 months ago
text_encoder openvino-model (#19) about 2 months ago
text_encoder_2 openvino-model (#19) about 2 months ago
tokenizer License, tags and diffusers updates (#2) about 2 months ago
vae_1_0 [Diffusers] Re-instate 0.9 VAE as default VAE … about 2 months ago
comparison.png 130 kB Upload 3 files about 2 months ago
model_index.json 609 Bytes Update model_index.json about 1 month ago
pipeline.png 80.2 kB Upload 3 files about 2 months ago
sd_xl_base_1.0.safetensors 6.94 GB LFS add weights about 2 months ago
sd_xl_base_1.0_0.9vae.sa… 6.94 GB LFS Upload sd_xl_base_1.0_0.9vae.safetensors about 2 months ago
sd_xl_offset_example-lor…
https://huggingface.co/stabilityai/stable-diffusion-xl-ba
Model na generovanie obrázkov tentokrát poskytuje **stabilityai** na HF

stabilityai/stable-diffusion-xl-r

huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0

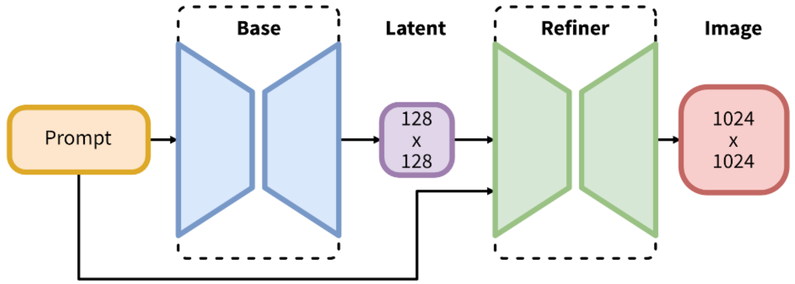
Hugging Face | Models | Datasets | Spaces | Docs | Solutions | Pricing | Log In | Sign Up

stabilityai / **stable-diffusion-xl-refiner-1.0** | like 849

Edit model card

SD-XL 1.0-refiner Model Card

Model



SDXL consists of an ensemble of experts pipeline for latent diffusion: In a first step, the base

Downloads last month
761,192

Hosted inference API

Text-to-Image

Your sentence here... | Compute

This model can be loaded on the Inference API on-demand.

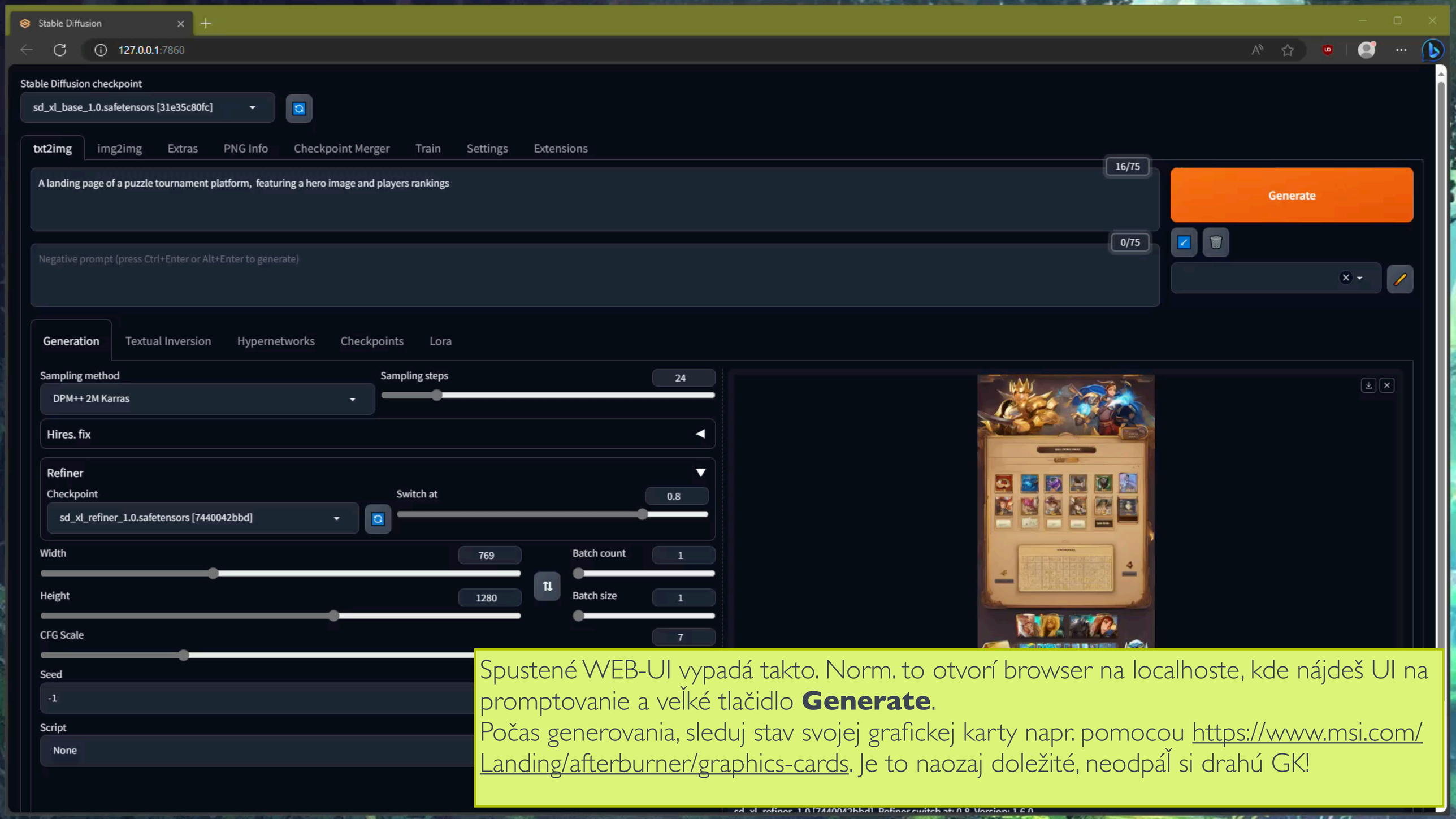
JSON Output | Maximize

Spaces using `stabilityai/stable-diffusion...` 93

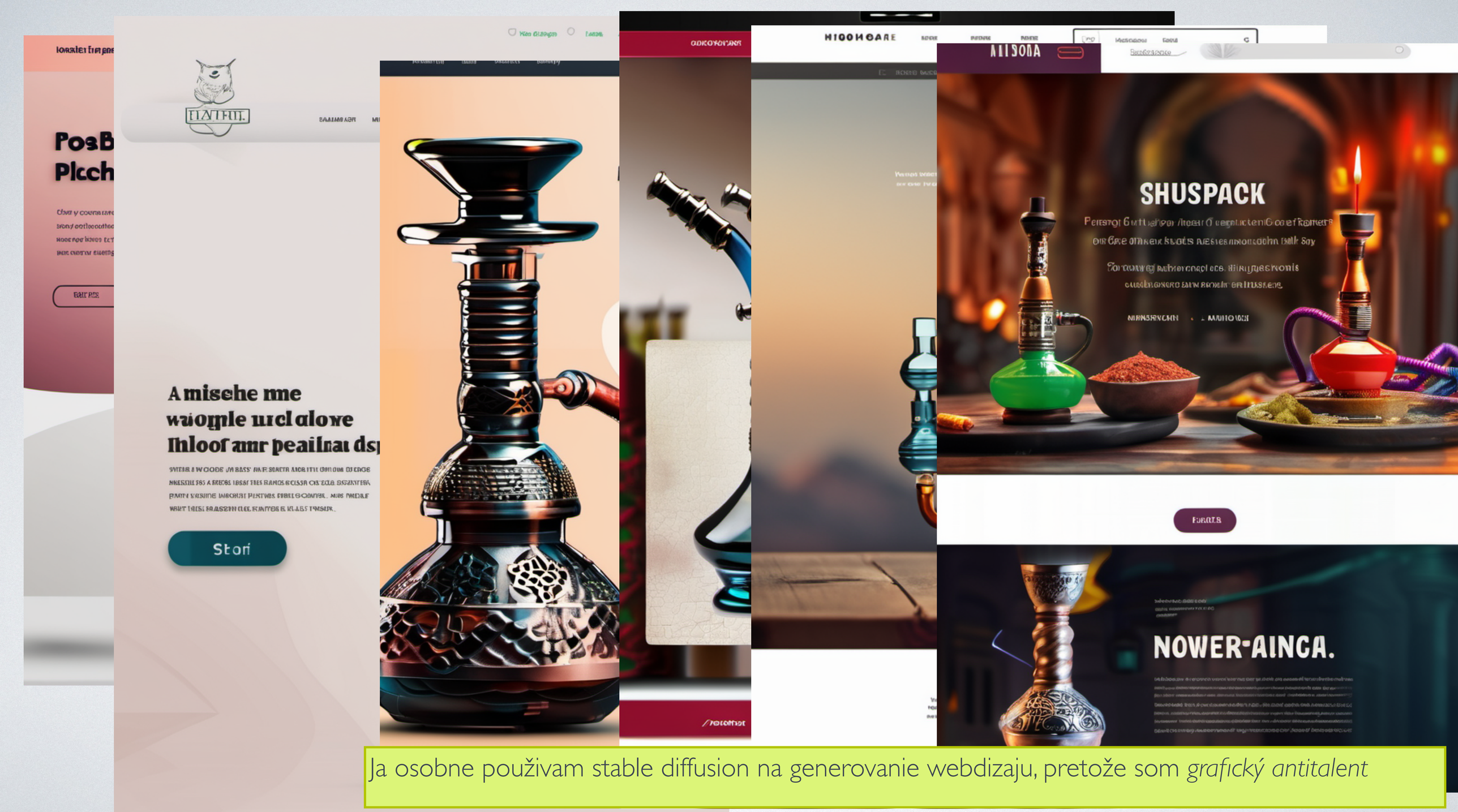
hysts/SD-XL | Manjushri/SDXL-1.0 | Manjushri/PhotoReal-V3.6 | Manjushri/SDXL-1.0-Img2Img-CPU | Manjushri/SDXL-1.0-CPU | jbilcke-hf/image-server-1 | pikto/Diffuser

Môžeš si stiahnúť aj tzv. ***refiner*** pre ešte lepšie výsledky.

https://huggingface.co/stabilityai/stable-diffusion-xl-ref



Spustené WEB-UI vypadá takto. Norm. to otvorí browser na localhoste, kde nájdeš UI na promptovanie a veľké tlačidlo **Generate**.
Počas generovania, sleduj stav svojej grafickej karty napr. pomocou https://www.msi.com/Landing/afterburner/graphics-cards. Je to naozaj doležité, neodpáľ si drahú GK!



Ja osobne používam stable diffusion na generovanie webdizaju, pretože som *grafický antitalent*

Aj logá to generuje veľmi pekné

EOF

twitter.com/srigi