



# How-To AI

locally & offline

**quantization**

**lora**

**refinement**

**fine-tuning**

**llama**

**Hugging Face**

**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 rozbehat' 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**)

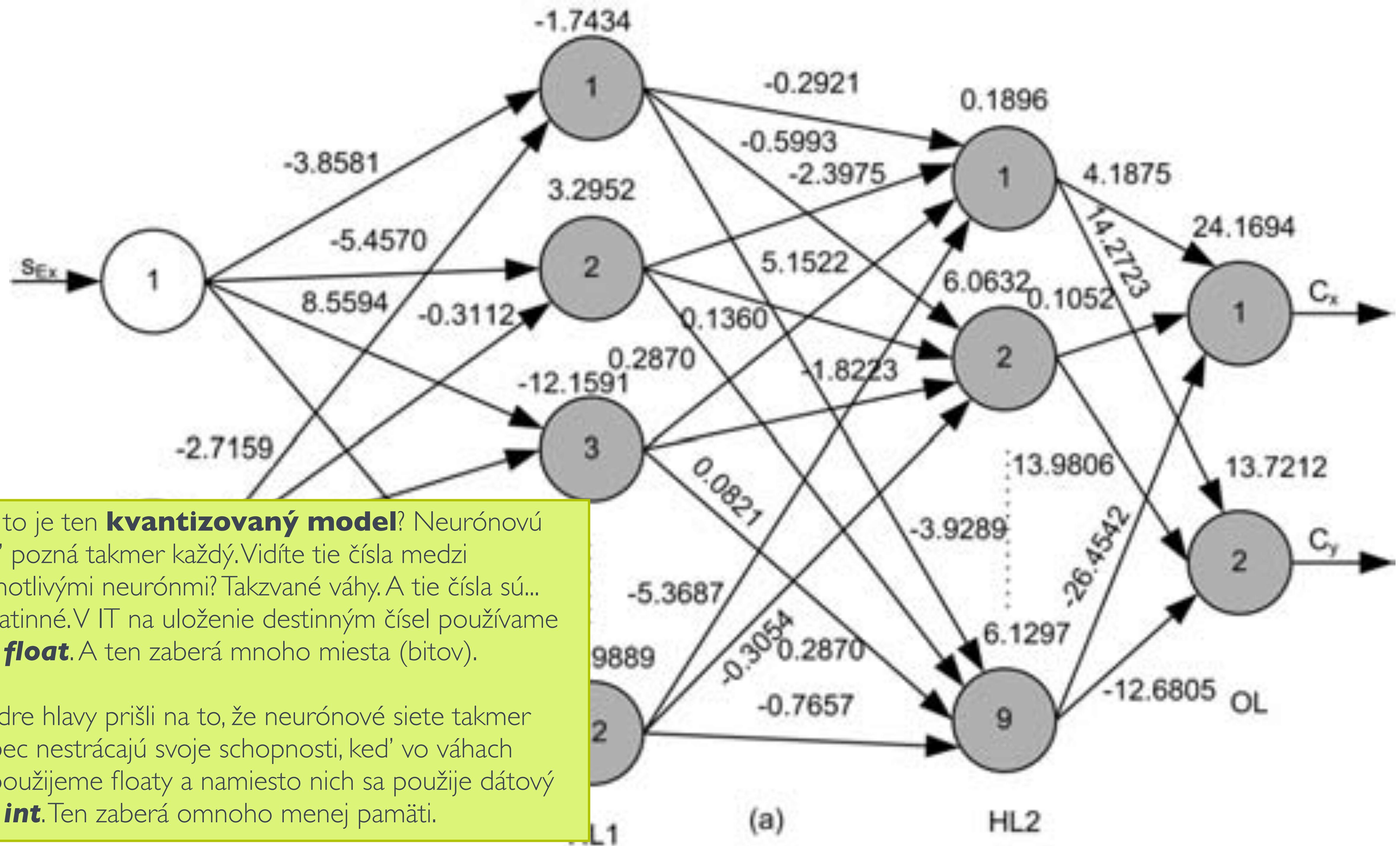
...ale hlavne pre tieto dôvody.

# **2 GROUPS OF GENERATIVE AIs**

- 1. LLMs (CHATGPT)**
- 2. IMAGE GENERATION (MIDJOURNEY)**

# HOW-TO START WITH **LLMS**

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





# Hugging Face

**NEW** Create Assistants in HuggingChat

## The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

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

The screenshot shows the Hugging Face website's search interface. At the top, there are tabs for Tasks, Libraries, Datasets, Languages, Licenses, and Other. Below the search bar, there are sections for Multimodal tasks (Text-to-Image, Image-to-Text, Text-to-Video, Visual Question Answering, Document Question Answering, Graph Machine Learning), Computer Vision tasks (Depth Estimation, Image Classification, Object Detection, Image Segmentation, Image-to-Image, Unconditional Image Generation, Video Classification, Zero-Shot Image Classification), Natural Language Processing tasks (Text Classification, Token Classification, Table Question Answering, Question Answering, Zero-Shot Classification, Translation, Summarization, Conversational, Text Generation, Text2Text Generation, Sentence Similarity), Audio tasks (Text-to-Speech, Automatic Speech Recognition, Audio-to-Audio, Audio Classification, Voice Activity Detection), Tabular tasks (Tabular Classification, Tabular Regression), Reinforcement Learning tasks (Reinforcement Learning, Robotics), and Robotics. On the right, a list of models is displayed, each with a name, description, update time, and statistics (e.g., 25.2k, 64, 393, 1.3k, 136, 1.87k, 2.66k, 334, 328, 64, 288k, 899, 12.5k, 332, 448k, 5.72k, 782k, 2.81k, 6.18k, 57).

Model	Description	Updated	Statistics
meta-llama/Llama-2-70b	Text Generation • Updated 4 days ago	25.2k	64
stabilityai/stable-diffusion-xl-base-0.9	Updated 6 days ago	2.01k	393
openchat/openchat	Text Generation • Updated 2 days ago	1.3k	136
lillyasvieu/ControlNet-v1-1	Updated Apr 26	1.87k	
cerspense/zeroscope_v2_XL	Updated 3 days ago	2.66k	334
meta-llama/Llama-2-13b	Text Generation • Updated 4 days ago	328	64
tiiuae/falcon-40b-instruct	Text Generation • Updated 27 days ago	288k	899
WizardLM/WizardCoder-15B-V1.0	Text Generation • Updated 3 days ago	12.5k	332
CompVis/stable-diffusion-v1-4	Text-to-Image • Updated about 17 hours ago	448k	5.72k
stabilityai/stable-diffusion-2-1	Text-to-Image • Updated about 17 hours ago	782k	2.81k
Salesforce/xgen-7b-8k-inst	Text Generation • Updated 4 days ago	6.18k	57

codellama/CodeLlama-70b-Instruct-hf

like 187

Model card

Files and versions

Community 24

:

Train

Deploy

Use in Transformers

## Code Llama

Code Llama is a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. This is the repository for the 70B instruct-tuned version in the Hugging Face Transformers format. This model is designed for general code synthesis and understanding. Links to other models can be found in the index at the bottom.

Base Model	Python	Instruct
7B <a href="#">codellama/CodeLlama-7b-hf</a>	<a href="#">codellama/CodeLlama-7b-Python-hf</a>	<a href="#">codellama/CodeLlama-7b-Instruct-hf</a>
13B <a href="#">codellama/CodeLlama-13b-hf</a>	<a href="#">codellama/CodeLlama-13b-Python-hf</a>	<a href="#">codellama/CodeLlama-13b-Instruct-hf</a>
34B <a href="#">codellama/CodeLlama-34b-hf</a>	<a href="#">codellama/CodeLlama-34b-Python-hf</a>	<a href="#">codellama/CodeLlama-34b-Instruct-hf</a>
70B <a href="#">codellama/CodeLlama-70b-hf</a>	<a href="#">codellama/CodeLlama-70b-Python-hf</a>	<a href="#">codellama/CodeLlama-70b-Instruct-hf</a>

Edit model card

Downloads last month  
10,822



Safetensors Model size 69B params Tensor type BF16

## Inference API

Text Generation

Examples

Input a message to start chatting with codellama/CodeLlama-70b-Instruct-hf.

You are a helpful and honest code assistant expert in JavaScript. Please, provide all answers to programming questions in JavaScript

Write a function that computes the set of sums of all contiguous sublists of a given list.

Your sentence here...

Send

This model can be loaded on Inference API (serverless).

Model requires a Pro subscription; check out [hf.co/pricing](#) to learn more. Make sure to include your HF token in your query.

JSON Output

Maximize

Spaces using codellama/CodeLlama-70b-Instruct-hf 17

bigcode/bigcode-models-leaderboard Omnibus/Chatbot-Compare

Statical/STC-LLM bardsai/performance-lm-board

klavyelibey/codellama-CodeLlama-70b-Instruct-hf Omnibus/AI-book

AilexGPT/Chatbot-Compare ColamanAI/hf-lm-api

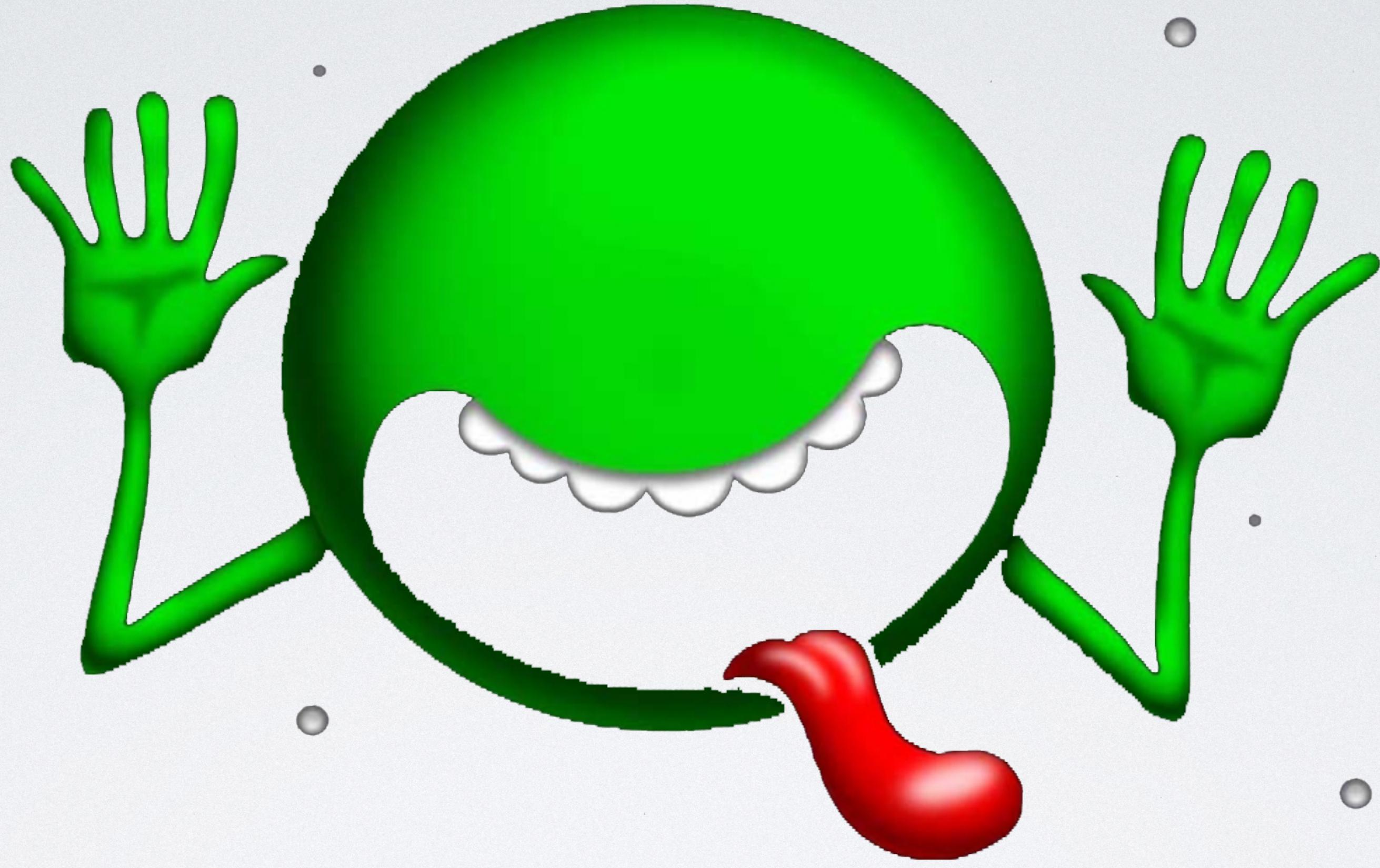
INDHU123/Chatbot\_comparing\_results

An-Egoistic-Developer-Full-Of-Knowledge/codellama-CodeLlama-70b-Instruct-hf

Nojo, tak som si našiel, s ktorým si vytvorím vlastného **copilota**.

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

To nedám 😢



DON'T PANIC



Tom Jobbins

PRO

TheBloke

Follow

17769 followers · 16 following

X TheBlokeAI

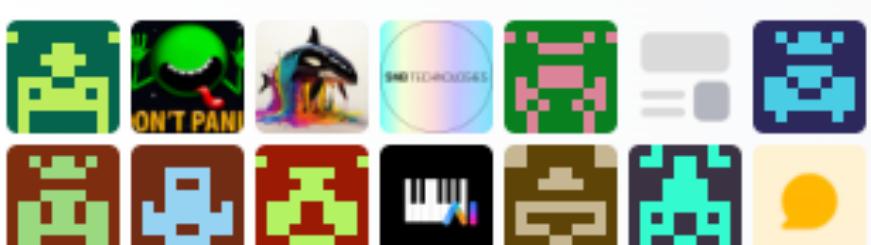
#### AI & ML interests

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

TheBloke/CapybaraHermes-2.5-Mistral-7B-GPTQ

Updated Jan 31 · 3.68k · 31

TheBloke/CapybaraHermes-2.5-Mistral-7B-GGUF

Updated Jan 31 · 162 · 50

TheBloke/KafkaLM-70B-German-V0.1-AWQ

Text Generation · Updated Jan 31 · 24 · 2

TheBloke/CodeLlama-70B-Python-GPTQ

Text Generation · Updated Jan 31 · 228 · 15

TheBloke/CodeLlama-70B-Python-GGUF

Text Generation · Updated Jan 31 · 34 · 29

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

▼ Expand 3863 models

 TheBloke/Samantha-1.11-CodeLlama-34B-GPTQ

Text Generation • Updated 2 days ago • ↓ 174 • ❤ 19

 TheBloke/CodeLlama-34B-GPTQ

Text Generation • Updated 2 days ago • ↓ 561 • ❤ 15

**GGML**



• old format, not supported by LLAMA.cpp

 THEBLOKE/CodeLlama-34B-Instruct-GPTQ

Text Generation • Updated 2 days ago • ↓ 7.47k • ❤ 36

 TheBloke/Samantha-1.11-

Updated 2 days ago • ↓ 4 • ❤ 5

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

 TheBloke/Samantha-1.11-70B-GPTQ

Text Generation • Updated 2 days ago • ↓ 54 • ❤ 3

**GPTQ**



• for GPU acceleration (multi-quantization)

 TheBloke/CodeLlama-13B-Python-GPTQ

**GGUF** • for METAL acceleration ( Mac)

 TheBloke/CodeLlama-7B-Python-GPTQ

Text Generation • Updated 2 days ago • ↓ 1.9k • ❤ 21

 TheBloke/CodeLlama-7B-GPTQ

 TheBloke/CodeLlama-34B-Instruct-GGUF

Text Generation • Updated 2 days ago • ↓ 143 • ❤ 63

**AWQ**



• mem. efficient GPU acceleration (nVidia only)

 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

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

## 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



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 \(a16z\)](#)

Edit model card

Downloads last month  
212



Hosted inference API (i)

Text Generation

Inference API has been turned off for this model.

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

mikeee/codellama-13b-instruct-gguf

## CodeLlama 13B Instruct - GGUF

- Model creator: [Meta](#)
- Original model: [CodeLlama 13B Instruct](#)

## Provided files

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

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

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

viz. iné kvantizácie...

Pozor, nemôžeš použiť všetkú pamäť svojho počítača!

# MEMORY BUDGET GAME

**24GB**

Nejakú pamäť si ukúsne operačný systém...

# MEMORY BUDGET GAME

system

**22.5GB**

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

# MEMORY BUDGET GAME

system

Chrome

**20GB**

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

# MEMORY BUDGET GAME

system    Chrome    Webstorm    **18-16GB**

# CodeLlama-34B-Instruct-GGUF

Modely s veľkým počtom váh vyžadujú nakoľko veľké množstvo pamäti.

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

# HOW-TO START WITH LLMS

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](https://github.com/ggerganov/llama.cpp).

The screenshot shows the GitHub repository page for `llama.cpp`. The repository is public and has 466 issues, 101 pull requests, and 4 projects. It has 5.7k forks and 40.7k stars. The latest commit was made 10 hours ago. The repository description states: "Port of Facebook's LLaMA model in C/C++". The repository features a large banner with the text "LLaMA C++". Below the banner, there are links for "CI passing", "license MIT", "Roadmap", "Manifesto", and "ggml". A note says "Inference of LLaMA model in pure C/C++". On the right side, there are sections for "Contributors" (346) and a "Contributor pie chart" showing the distribution of code contributions by language: C (42.5%), C++ (27.5%), Cuda (11.9%), Python (6.8%), Metal (3.8%), Objective-C (2.9%), and Other (4.6%).



↳ b2548  
↳ e82f9e2 ✓  
Compare ▾

## b2548 Latest

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

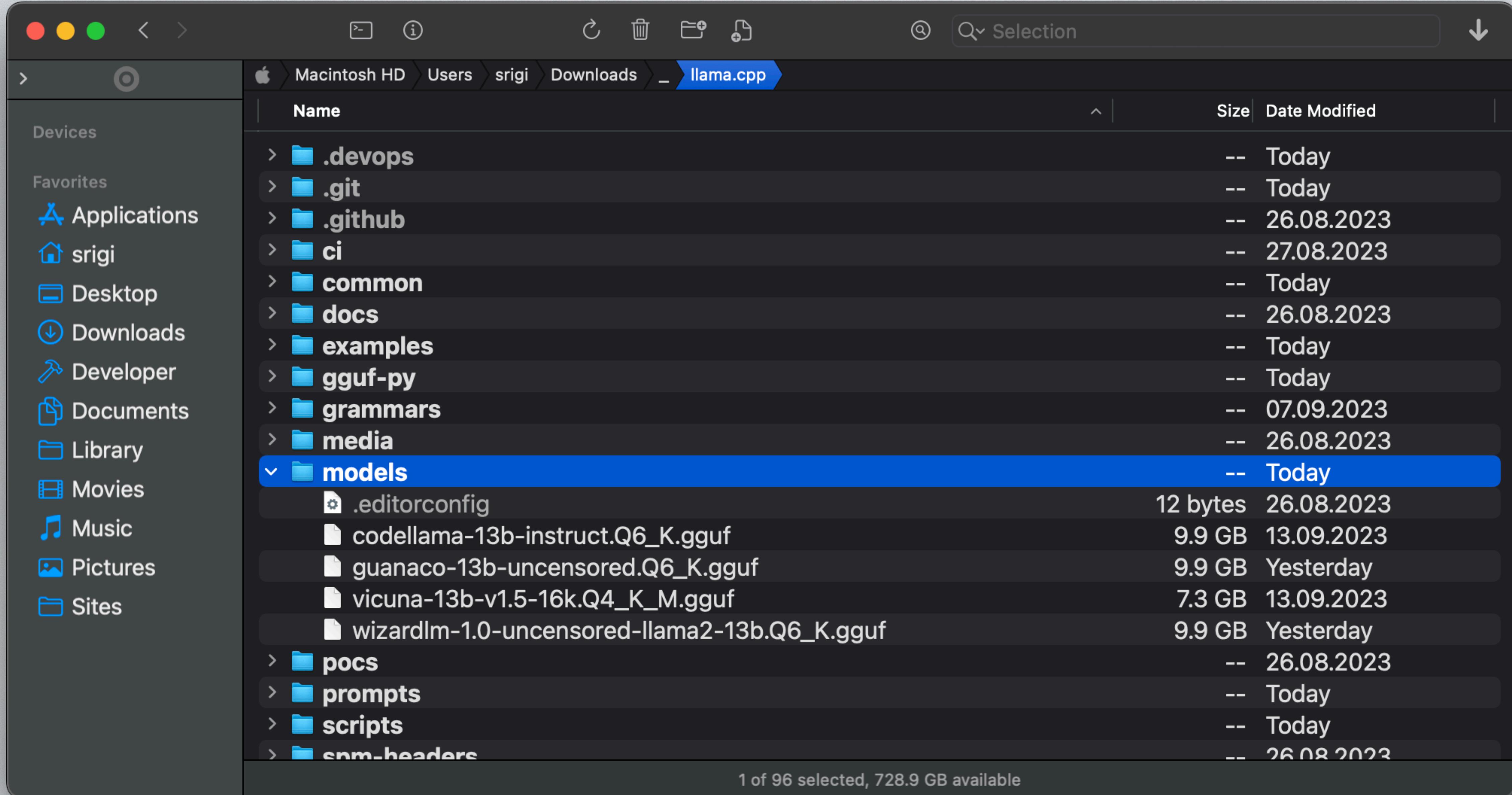
- \* Fix batched impl
- \* Maintain previous behaviour for igpu
- \* retrigger CI

### ▼ Assets

18

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

Netráp sa s nejakým komplilovaním. Na stránke **Releases** nájdeš predkomplilované binárky pre takmer každý HW a s podporou akcelerácie.



Llama.cpp má pre modely predpravenú zložku. Môžeš si tam hodit' 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

Na detail stránke modelu od týpka si ešte nájdi sekciu **Prompt template**. Budeš to potrebovať hned' v d'alšom kroku.

As this model is based on L

Ukazuje ako model očakáva, že mu zadáš prompt.

and the license files for that are additionally included. It should therefore be

```
#!/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 Windows nazývaný tiež bat'ák), ktorým spustíš llama.cpp. Na obrázku vidíš odladený script pre model llama2. Najdôležitejší je argument **n-gpu-layers**, ktorým určuješ kolko vrstiev neurónov pošleš na akceleráciu na GPU. Počet layerov modelu zistíš 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 \

TheBloke (Tom Jobbins) x

huggingface.co/TheBloke

uncensored | 1/61 | ^ v x

TheBloke

Text Generation • Updated 8 days ago • ↓ 4.87k • ❤ 52

TheBloke/Nous-Hermes-Llama2-GGML

Updated 8 days ago • ↓ 269 • ❤ 92

TheBloke/Nous-Hermes-Llama2-GGUF

Updated 8 days ago • ↓ 13 • ❤ 7

TheBloke/Guanaco-13B-Uncensored-GPTQ

Conversational • Updated 8 days ago • ↓ 40 • ❤ 1

TheBloke/Guanaco-7B-Uncensored-GPTQ

Conversational • Updated 8 days ago • ↓ 35 • ❤ 3

TheBloke/Zarablend-L2-7B-GPTQ

Text Generation • Updated 8 days ago • ↓ 893 • ❤ 9

TheBloke/Guanaco-13B-Uncensored-GGUF

Conversational • Updated 8 days ago • ↓ 22 • ❤ 5

TheBloke/LLama-2-PeanutButter\_v19\_R8-7B-GPTQ

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

# HOW-TO START WITH IMAGE GENERATION





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



Ideálne je mat' k dispozícii takúto hračku

AUTOMATIC1111/stable-diffusion-webui

github.com/AUTOMATIC1111/stable-diffusion-webui

stable-diffusion-webui Public

Watch 865 Fork 20.4k Star 102k

master Branches Tags

AUTOMATIC1111 Merge branch 'release\_ca...' 2 weeks ago 5,986

.github update bug report template to include sys... 2 weeks ago

configs disable EMA weights for instructpix2pix m... 8 months ago

embeddings add embeddings dir last year

extensions-builtin Merge pull request #12838 from bluelove... 2 weeks ago

extensions delete the submodule dir (why do you kee... last year

html hide cards for networks of incompatible s... 2 months ago

javascript get progressbar to display correctly in ext... 2 weeks ago

localizations Remove old localizations from the main re... 10 months ago

models Add support for the Variations models (un... 6 months ago

modules Merge pull request #12838 from bluelove... 2 weeks ago

Stable Diffusion web UI

web ai deep-learning torch

pytorch unstable

image-generation gradio

diffusion upscaling text2image

image2image img2img ai-art

txt2img stable-diffusion

AGPL-3.0 license

Cite this repository

Activity

102k stars

865 watching

20.4k forks

1.6.0 Latest

2 weeks ago

Ako runtime použí [github.com/AUTOMATIC1111/stable-diffusion-webui](https://github.com/AUTOMATIC1111/stable-diffusion-webui)

AUTOMATIC1111/stable-diffusion-webui

github.com/AUTOMATIC1111/stable-diffusion-webui

launch.py	add --dump-sysinfo, a cmd arg to dump li...	2 weeks ago
package.json	Add basic ESLint configuration for format...	4 months ago
pyproject.toml	Overhaul tests to use py.test	4 months ago
requirements-test.t...	Overhaul tests to use py.test	4 months ago
requirements.txt	update gradio to 3.41.2	3 weeks ago
requirements_versi...	update gradio to 3.41.2	3 weeks ago
screenshot.png	new screenshot	8 months ago
script.js	Frontend: only look at top-level tabs, not ...	4 months ago
style.css	get progressbar to display correctly in ext...	2 weeks ago
webui-macos-env.sh	Mac k-diffusion workarounds are no long...	last month
webui-user.bat	revert change to webui-user.bat	10 months ago
webui-user.sh	Vendor in the single module used from ta...	4 months ago
webui.bat	Restart: only do restart if running via the ...	3 months ago
webui.py	Mer...	
webui.sh	Mer...	

README.md

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 klude.

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

s. stabilityai/stable-diffusion-xl-base-1.0 like 2.63k

Model card Files Community 87

⋮ Deploy Use in Diffusers

patrickvonplaten	PipPoweraker	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...					

Model na generovanie obrázkov tentokrát poskytuje **stabilityai** na HF

<https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0/tree/main>

stabilityai/stable-diffusion-xl-refiner-1.0

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

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

## s. stabilityai/stable-diffusion-xl-refiner-1.0

like 849

### SD-XL 1.0-refiner Model Card

Edit model card

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...

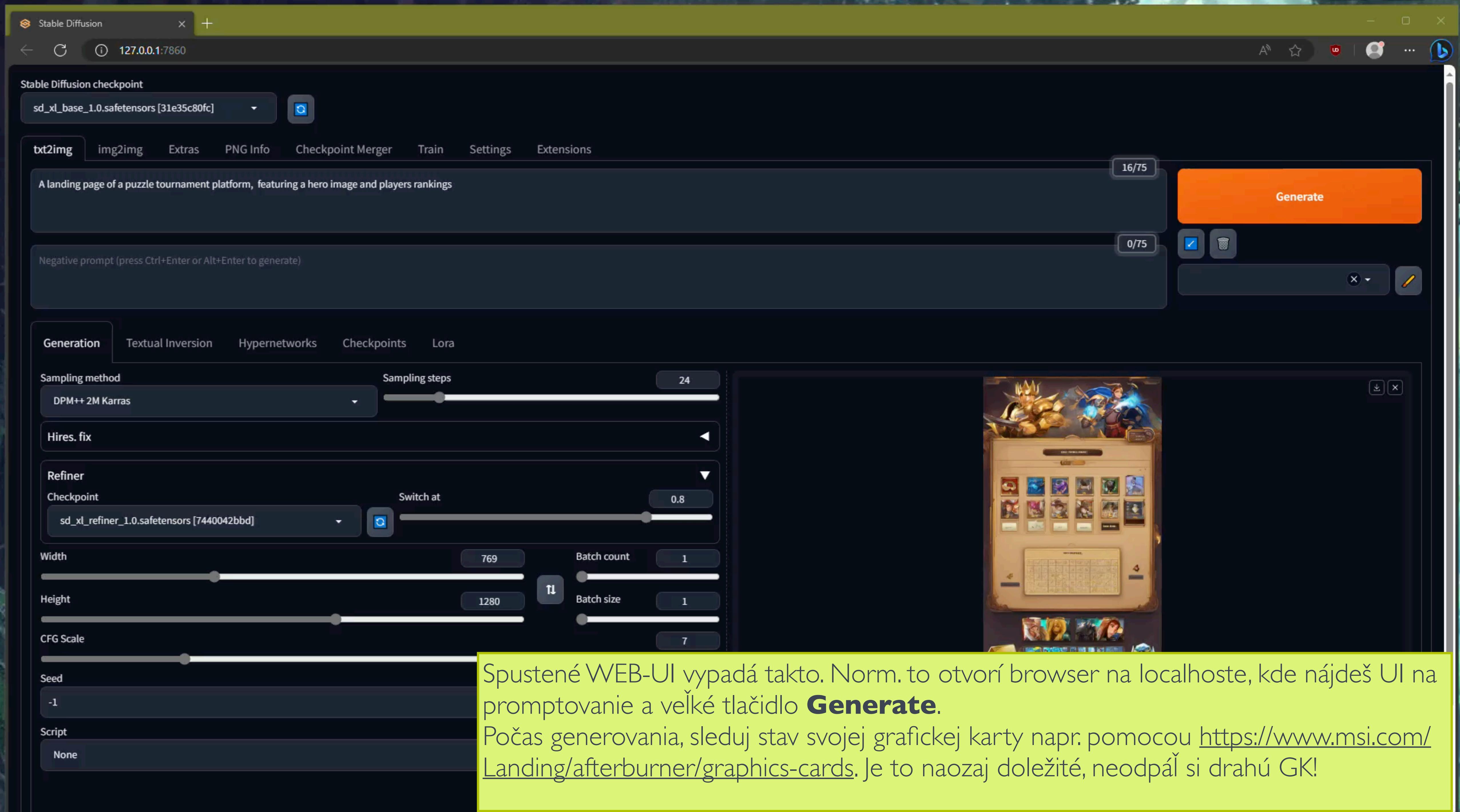
- hysts/SD-XL
- Manjushri/SDXL-1.0
- Manjushri/PhotoReal-V3.6
- Manjushri/SDXL-1.0-Img2Img-CPU
- Manjushri/SDXL-1.0-CPU
- jbilcke-hf/image-server-1
- pikto/Diffuser

**Model**

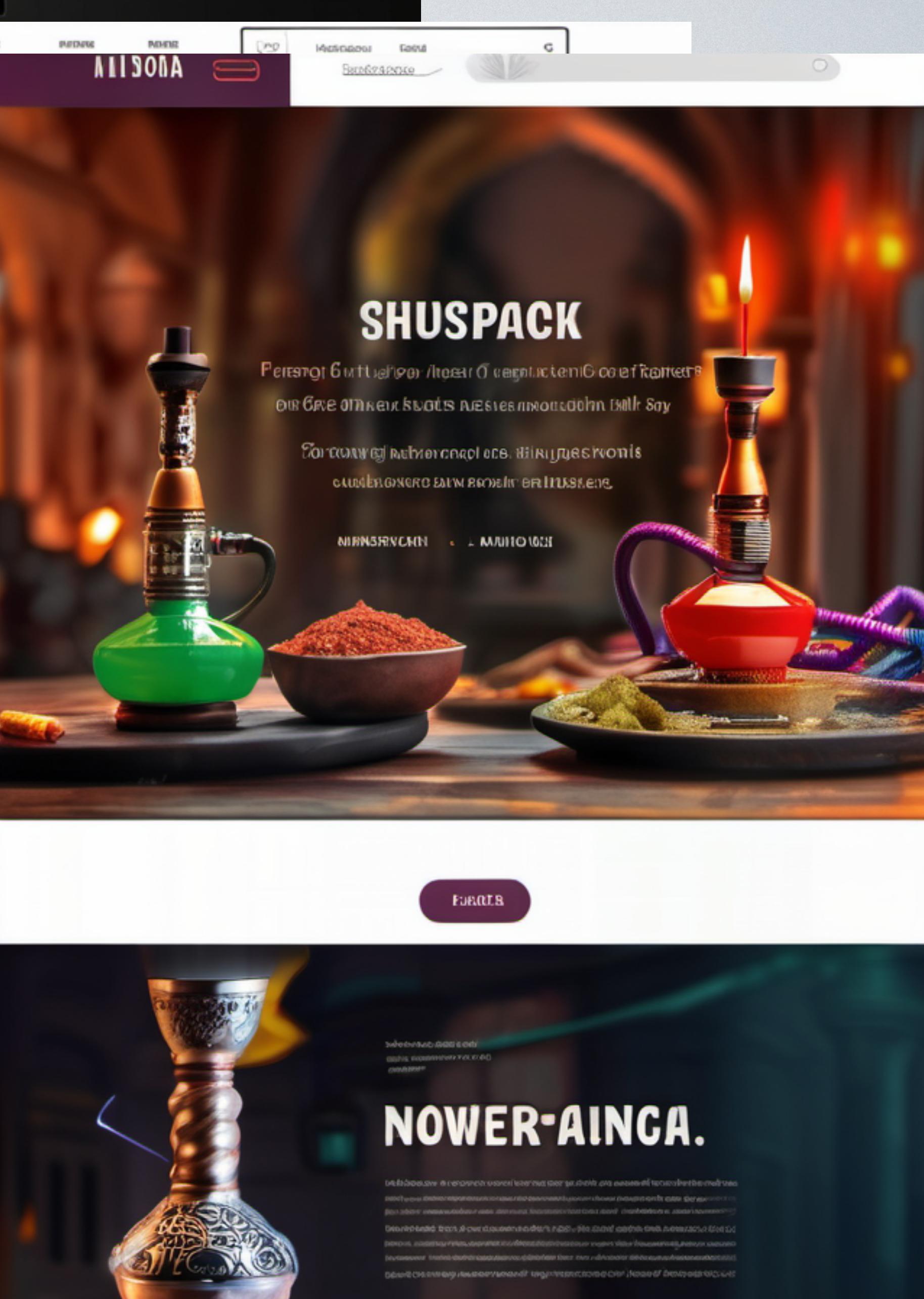
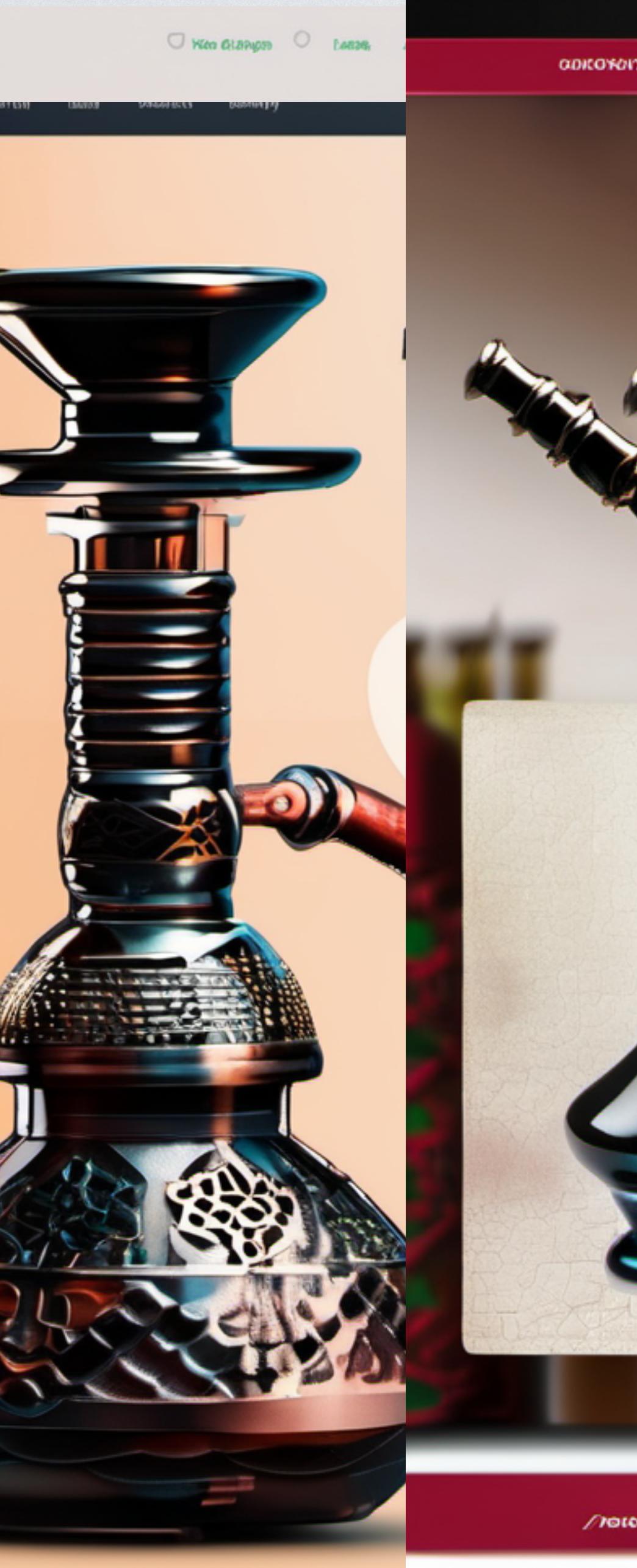
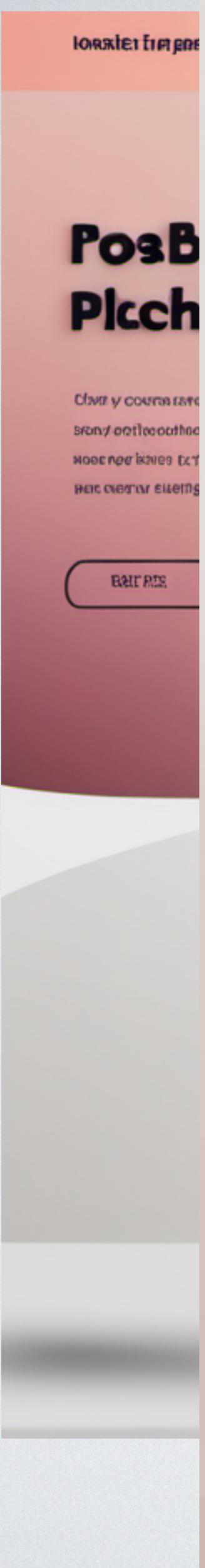
SDXL consists of an ensemble of experts pipeline for latent diffusion: In a first step, the base

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

<https://huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0>



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, neodpáť si drahú GK!

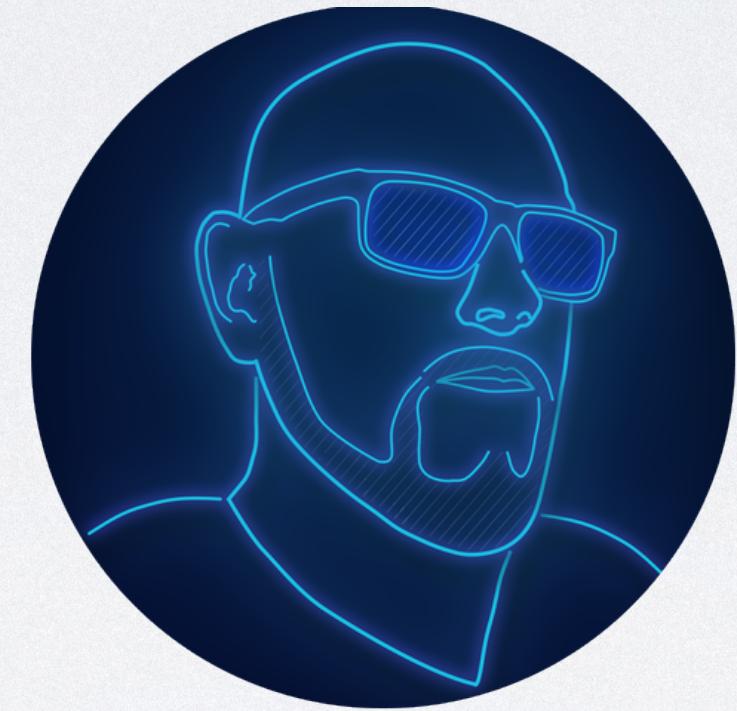


Ja osobne používam stable diffusion na generovanie webdizaju, pretože som grafický antitalent



Aj logá to generuje veľmi pekné

EOF



[twitter.com/srigi](https://twitter.com/srigi)