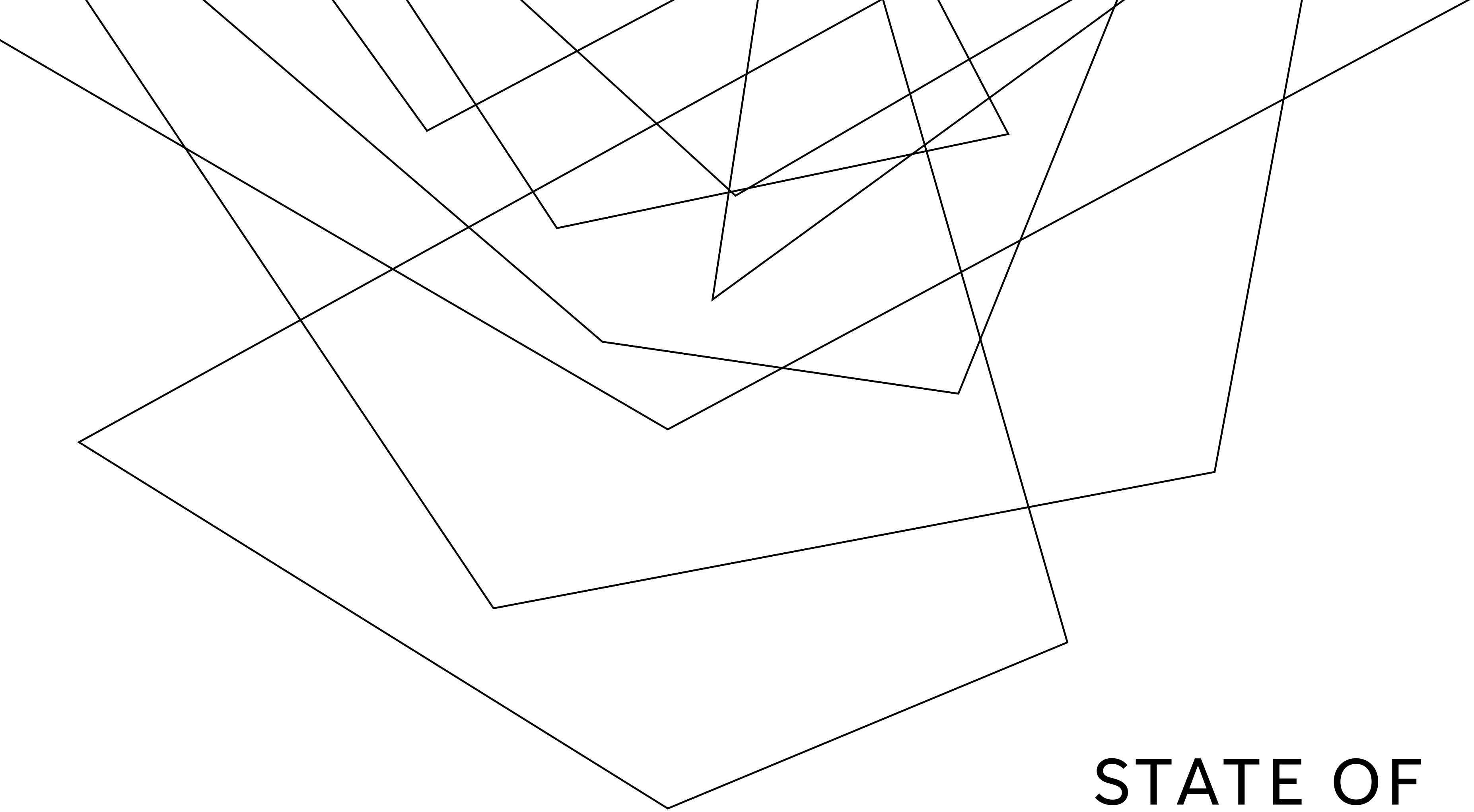


STATE OF GENERATIVE AI

November 2024 – Jakob Demler

WHO AM I? WHAT IS THIS?

software engineer by training

lost the joy of writing software ~2019

regained the joy of writing software with Github Copilot (~September 2022)

love playing around and working with AI

Nowadays I mostly write software in English

Biased, high-level overview of the space of generative AI I am familiar with

Aim to inspire or spark curiosity and give pointers

No theoretical background knowledge (I don't have that)

Optimize for fun!

(always feel free to interrupt and ask questions)

Follow here: https://github.com/JDemler/state-of-gen-ai

AGENDA

Introduction

Image Generation

Hands On

Text Generation

Audio Generation

Video Generation

Hands On

WHAT TOOLS DO I USE DAY TO DAY?

- **Software Development: cursor.com**
- **Search: perplexity.ai**
- **ChatGPT 4o with Canvas for text production**
- **ChatGPT o1-preview for hard problems**
- **ChatGPT Advanced Voice Mode for language learning**
- **Claude.ai for little "applets"**
- **Image Generation with recraft.ai (since Tuesday)**

IMAGE GENERATION

TIMELINE

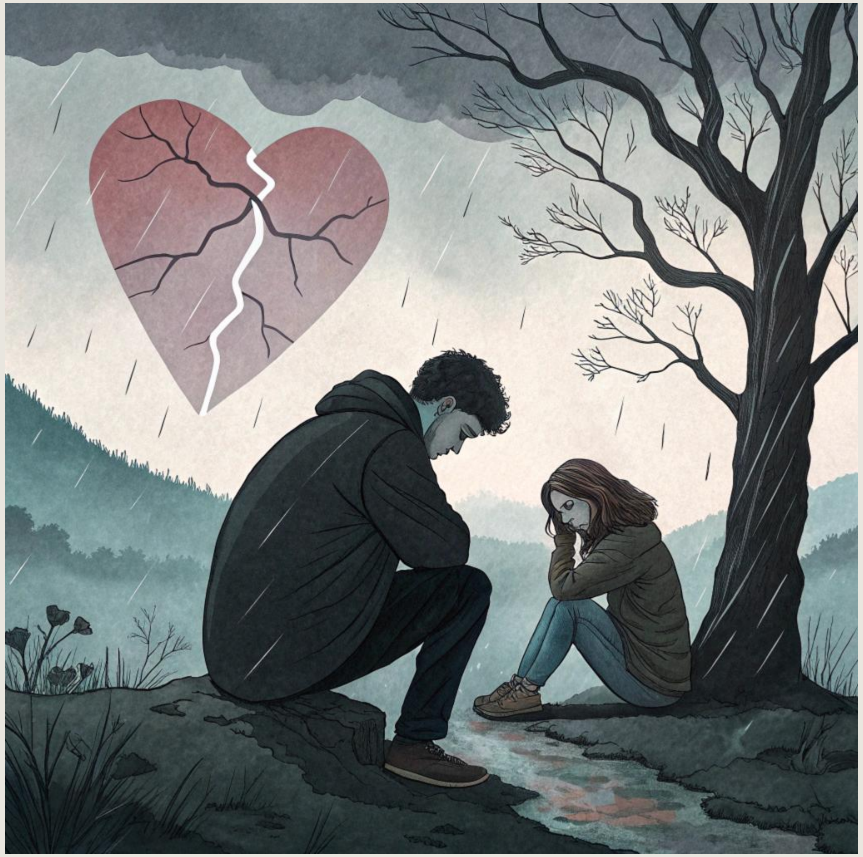
- First “usable” models:
- DALL-E 2 (April 2022)
- Midjourney (July 2022)
- Stable Diffusion(August 2022) (OpenSource)

- Current, state-of-the-art Models
- Flux 1.1 (October 2024, from Freiburg)
- Stable Diffusion 3.5 (October 2024)
- Midjourney v6 (July 2024)
- Recraft v3 (November 2024)

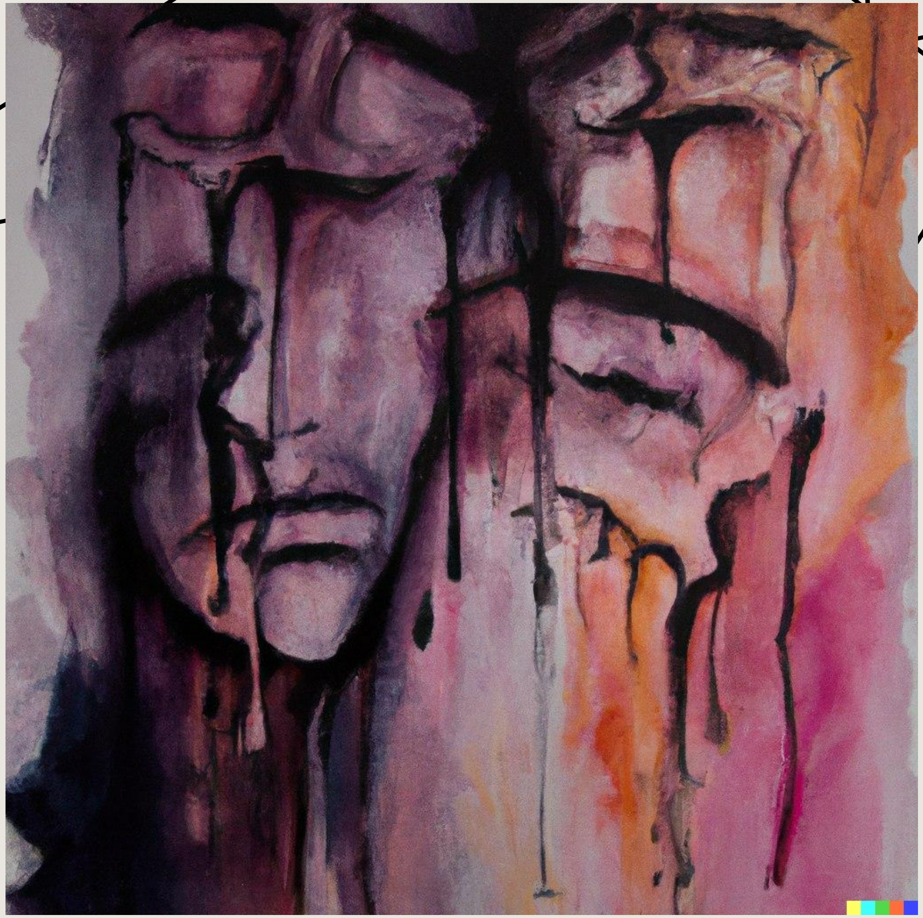
abstract oil on canvas painting of a kubernetes application drowning in complexity

Dalle 2 – 02-09-2022

RECRAFT V3 06-11-2024

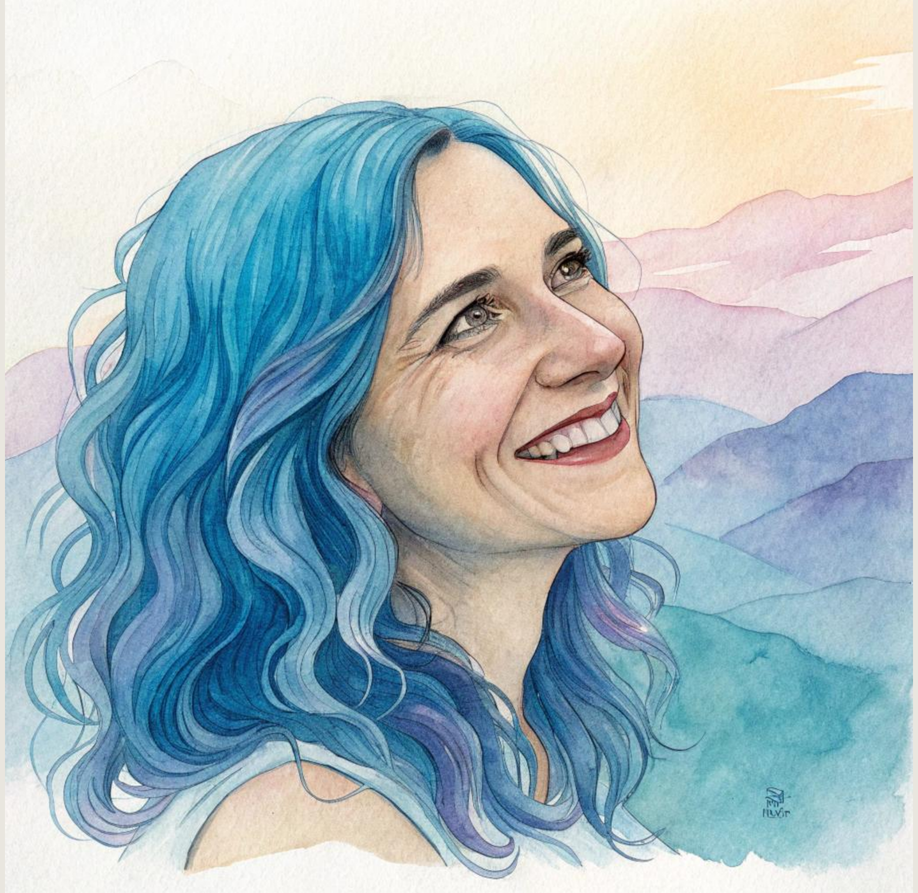


DALLE2 02-09-2022

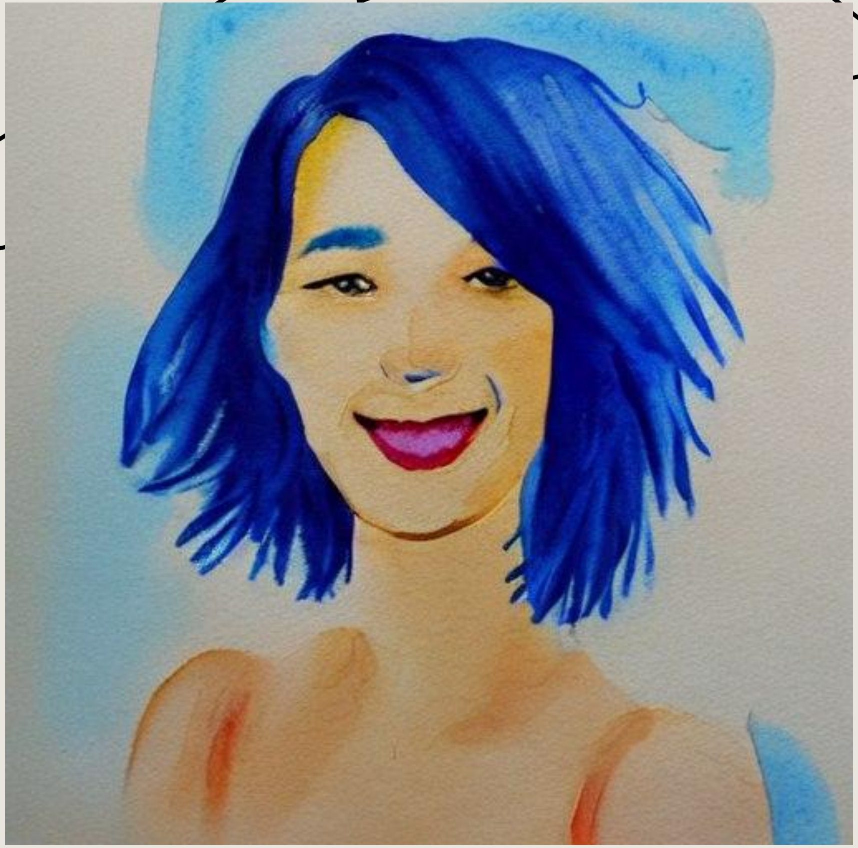


"watercolor painting of a broken love that cannot ever be fixed anymore, dark colors, hopelessness, deep sadness"

RECRAFT V3 06-11-2024



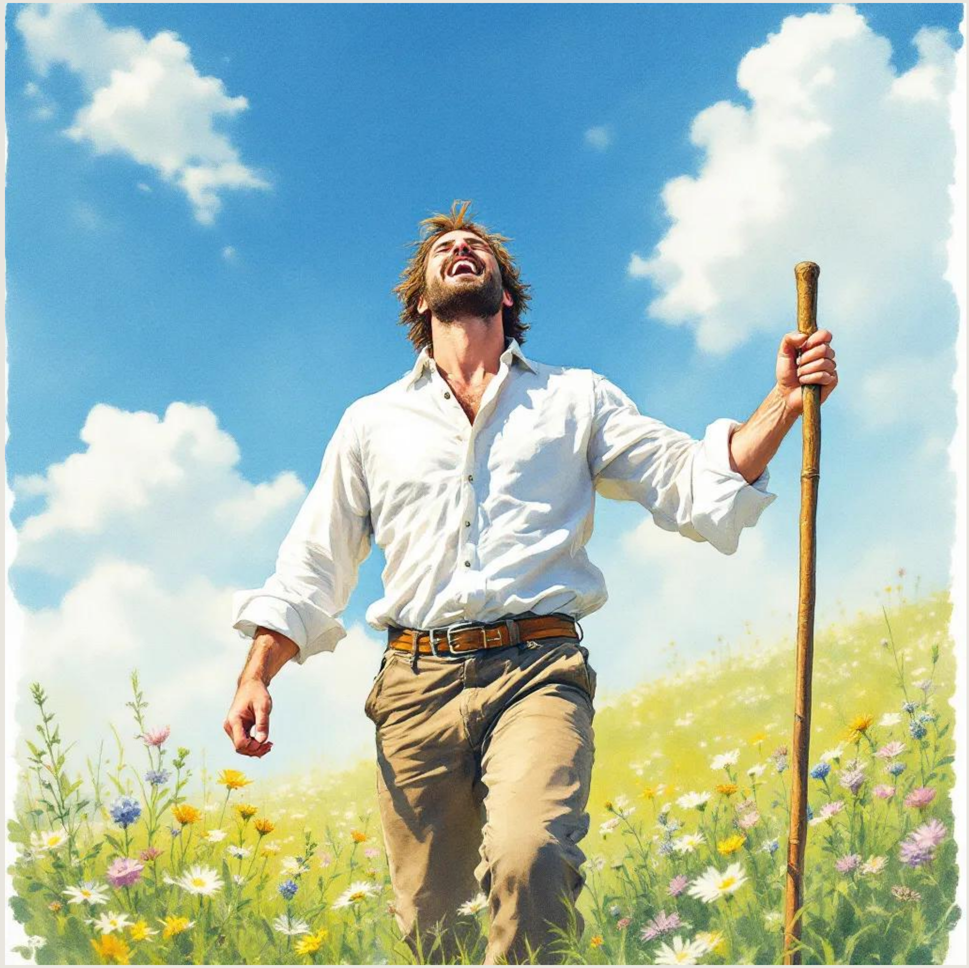
STABLE DIFFUSION 1.5 14-0 2022

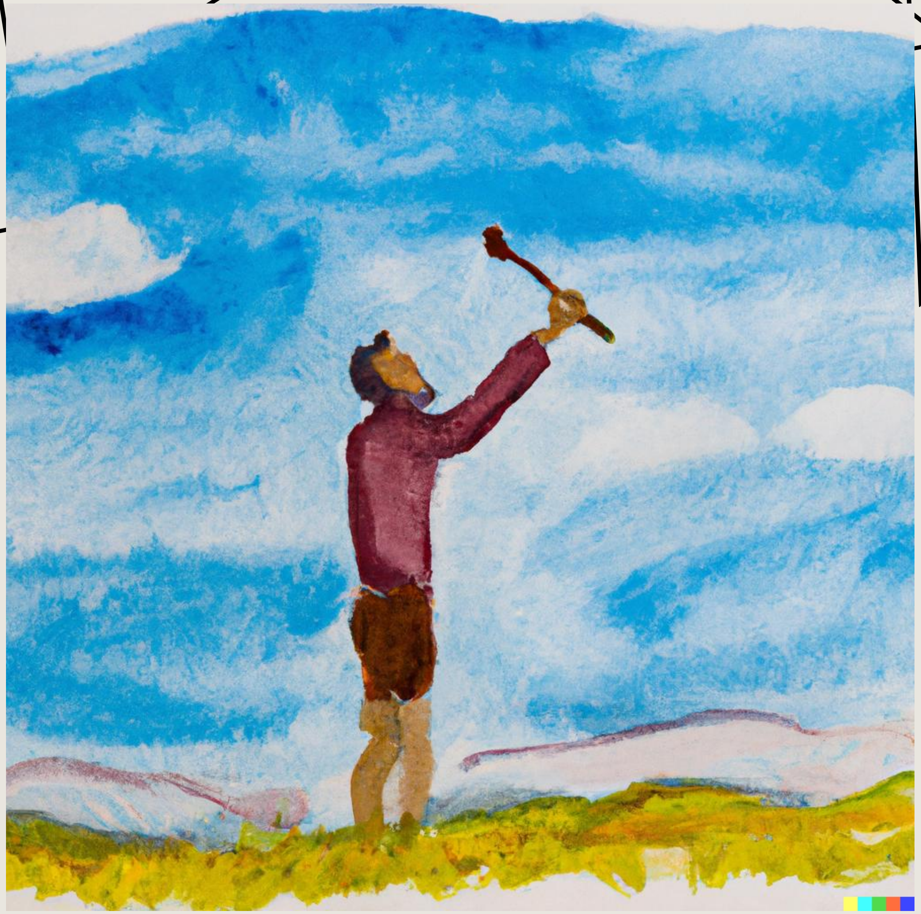


"watercolor painting of a smiling woman with blue hair"

FLUX 1.1 PRO 08-11-2024

DALLE-2 02-09-2022



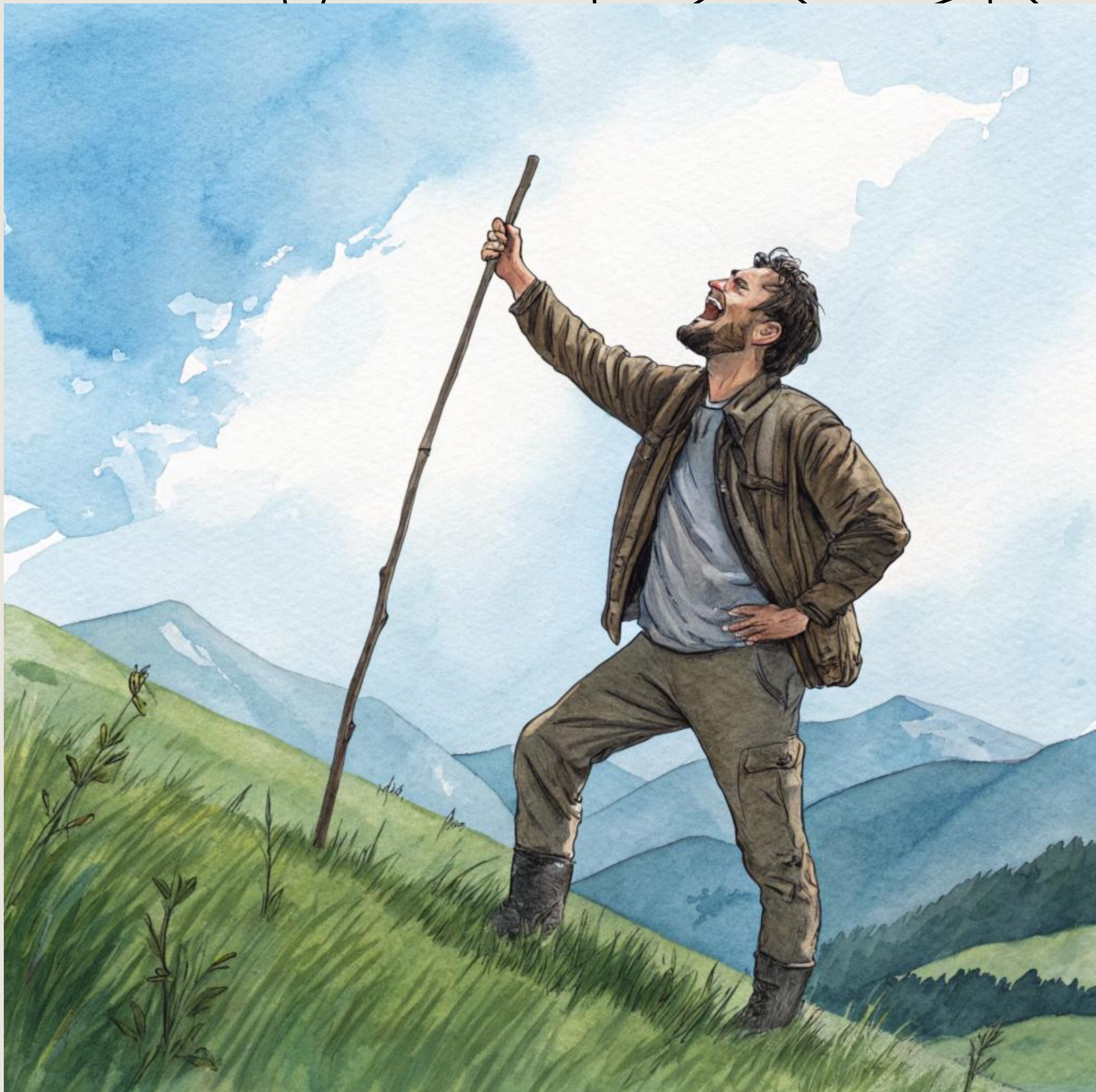


“watercolor painting of a man standing on a green alpine pasture with a stick in his hand, shouting of joy, happy, blue sky”

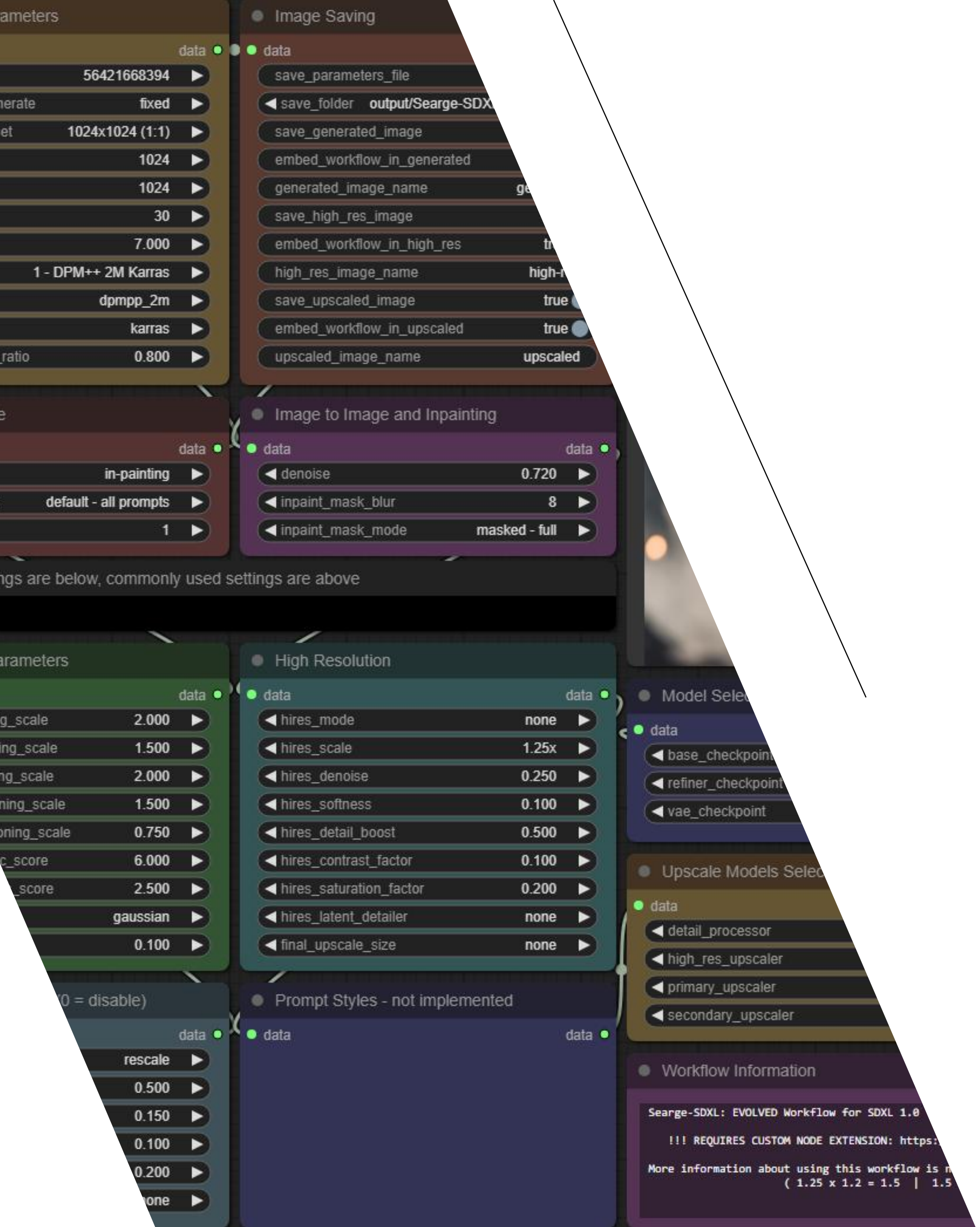
TRYING TO GET THE SPIRIT BACK INTO THE AI

Recraft v3 – 08-11-2024

- abstract rough watercolor painting of a handsome man, no details, thick lines, standing on a green alpine pasture with a stick in his hand, shouting of joy, happy, blue sky, artsy



WE HAVE COME FULL CIRCLE:



56421668394
nerate fixed
et 1024x1024 (1:1)
1024
1024
30
7.000
1 - DPM++ 2M Karras
dpmpp_2m
karras
ratio 0.800

save_parameters_file
save_folder output/Searge-SDX
save_generated_image
embed_workflow_in_generated
generated_image_name ge
save_high_res_image
embed_workflow_in_high_res tr
high_res_image_name high-r
save_upscaled_image true
embed_workflow_in_upscaled true
upscaled_image_name upscaled

in-painting
default - all prompts
1

denoise 0.720
inpaint_mask_blur 8
inpaint_mask_mode masked - full

ngs are below, commonly used settings are above

g_scale 2.000
ing_scale 1.500
ng_scale 2.000
ning_scale 1.500
oning_scale 0.750
c_score 6.000
_score 2.500
gaussian
0.100

hires_mode none
hires_scale 1.25x
hires_denoise 0.250
hires_softness 0.100
hires_detail_boost 0.500
hires_contrast_factor 0.100
hires_saturation_factor 0.200
hires_latent_detailer none
final_upscale_size none

base_checkpoin
refiner_checkpoint
vae_checkpoint

detail_processor
high_res_upscaler
primary_upscaler
secondary_upscaler

0 = disable)
rescale
0.500
0.150
0.100
0.200
one

Searge-SDXL: EVOLVED Workflow for SDXL 1.0
!!! REQUIRES CUSTOM NODE EXTENSION: https:
More information about using this workflow is
(1.25 x 1.2 = 1.5 | 1.5

CHARACTERISTICS

- can be run on consumer hardware in general
- Vibrant community of users (ControlNets, Finetunes, LORAs, Uis)
- Workflows can get very complicated

Communities:

- https://www.reddit.com/r/StableDiffusion/
- https://civitai.com/

Tools:

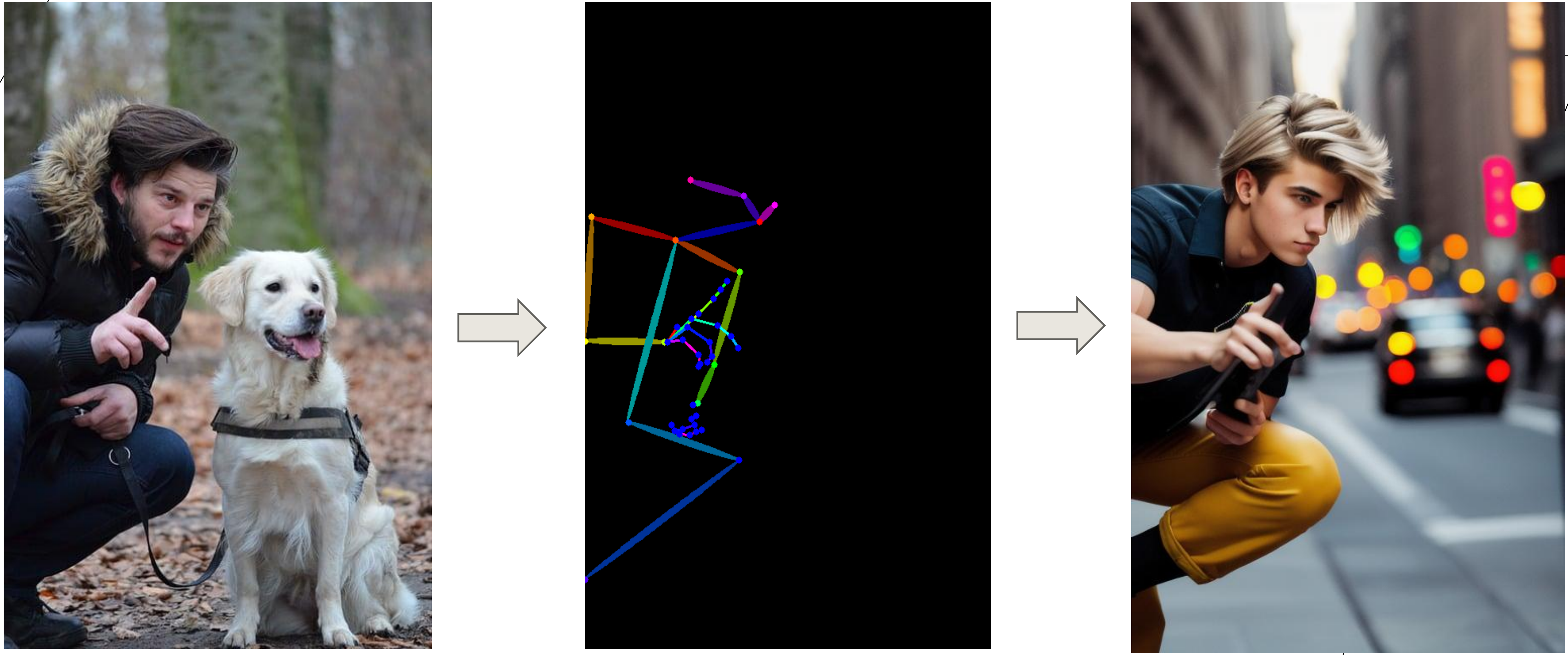
- https://github.com/comfyanonymous/ComfyUI
- https://github.com/AUTOMATIC1111/stable-diffusion-webui

DEMO

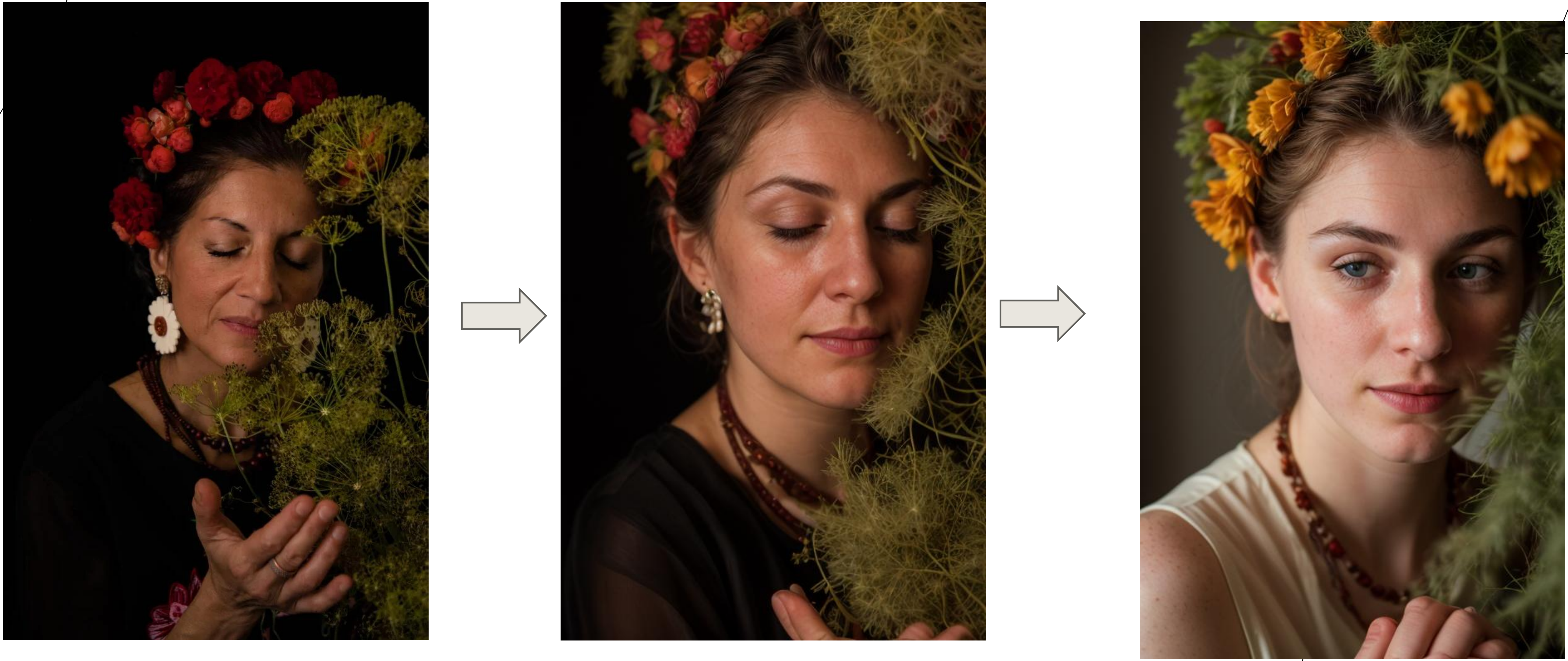
- Vast.ai (rented GPUs)
- ComfyUI
- Flux 1.1 “Schnell”

- Or hosted: replicate.com; recraft.ai

POSE



IP ADAPTER



TEXT GENERATION

TIMELINE

First “usable” models:

- ChatGPT (GPT 3.5) (December 2022)
- GPT-4 (April 2023)

Current, state-of-the-art models

- GPT-4o + o1-preview and mini (OpenAI)
- Claude Sonnet 3.5 (Anthropic)
- Google Gemini Pro (Google)
- Grok-2 (xAI)

Also: hundreds of smaller/opensource/finetune models

MetaAI: llama family of models (405b – 1b parameter, multimodal, open weights)

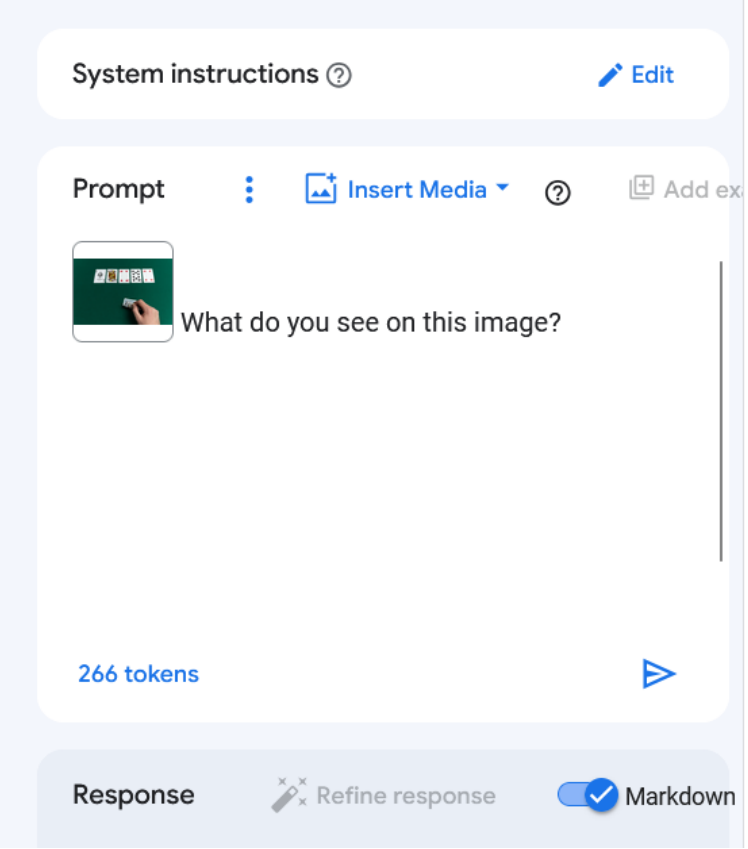
Alibaba: Qwen family of models (72b and 7b, multimodal)

Mistral, DeepSeek, probably a new one since I started talking

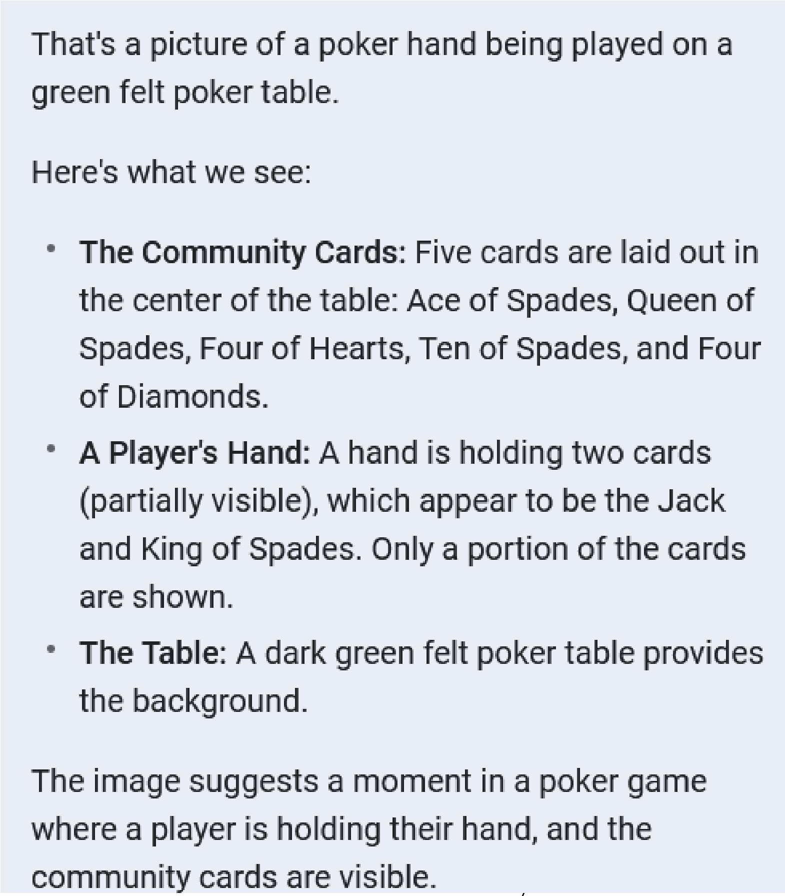
TIMELINE

Nowadays: Multimodality:

- **Not only Text but also Images, Video and Audio as input;**
- **In most cases still only text as output**



What do you see on this image?



That's a picture of a poker hand being played on a green felt poker table.

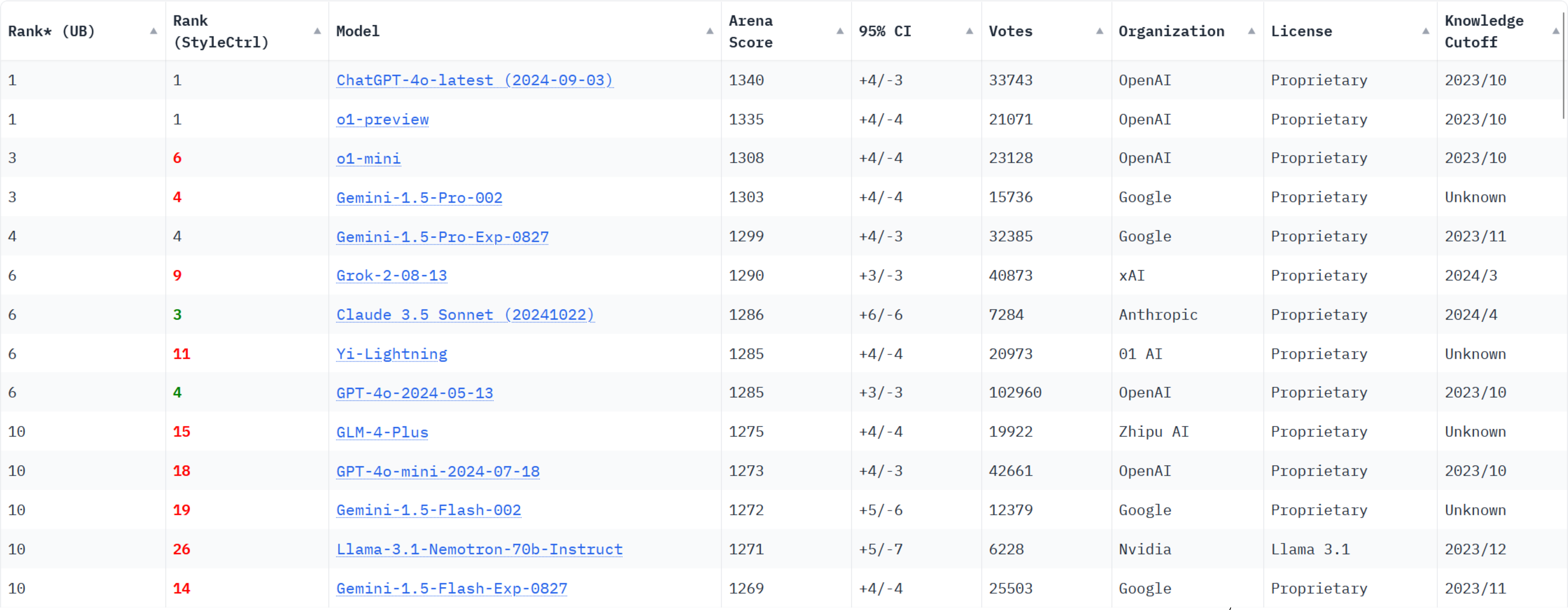
Here's what we see:

- **The Community Cards:** Five cards are laid out in the center of the table: Ace of Spades, Queen of Spades, Four of Hearts, Ten of Spades, and Four of Diamonds.
- **A Player's Hand:** A hand is holding two cards (partially visible), which appear to be the Jack and King of Spades. Only a portion of the cards are shown.
- **The Table:** A dark green felt poker table provides the background.

The image suggests a moment in a poker game where a player is holding their hand, and the community cards are visible.

HOW CAN I GET AN OVERVIEW?

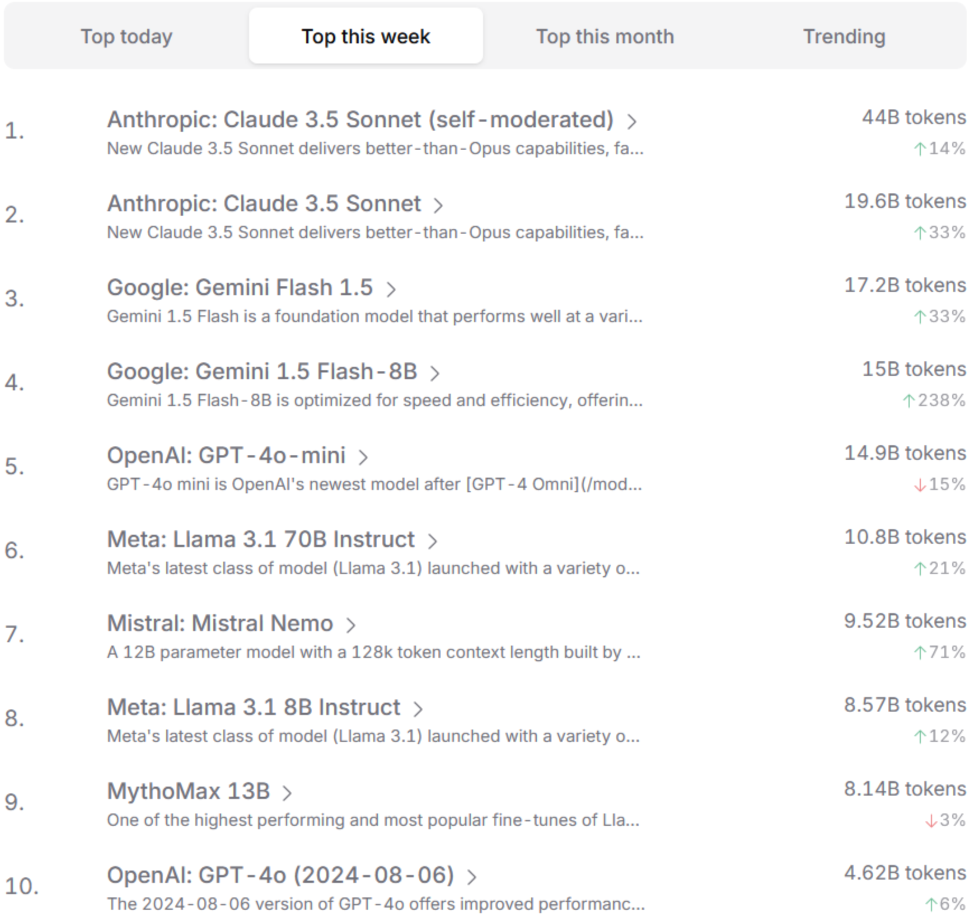
https://lmarena.ai/?leaderboard (LLMs ranked by ELO)



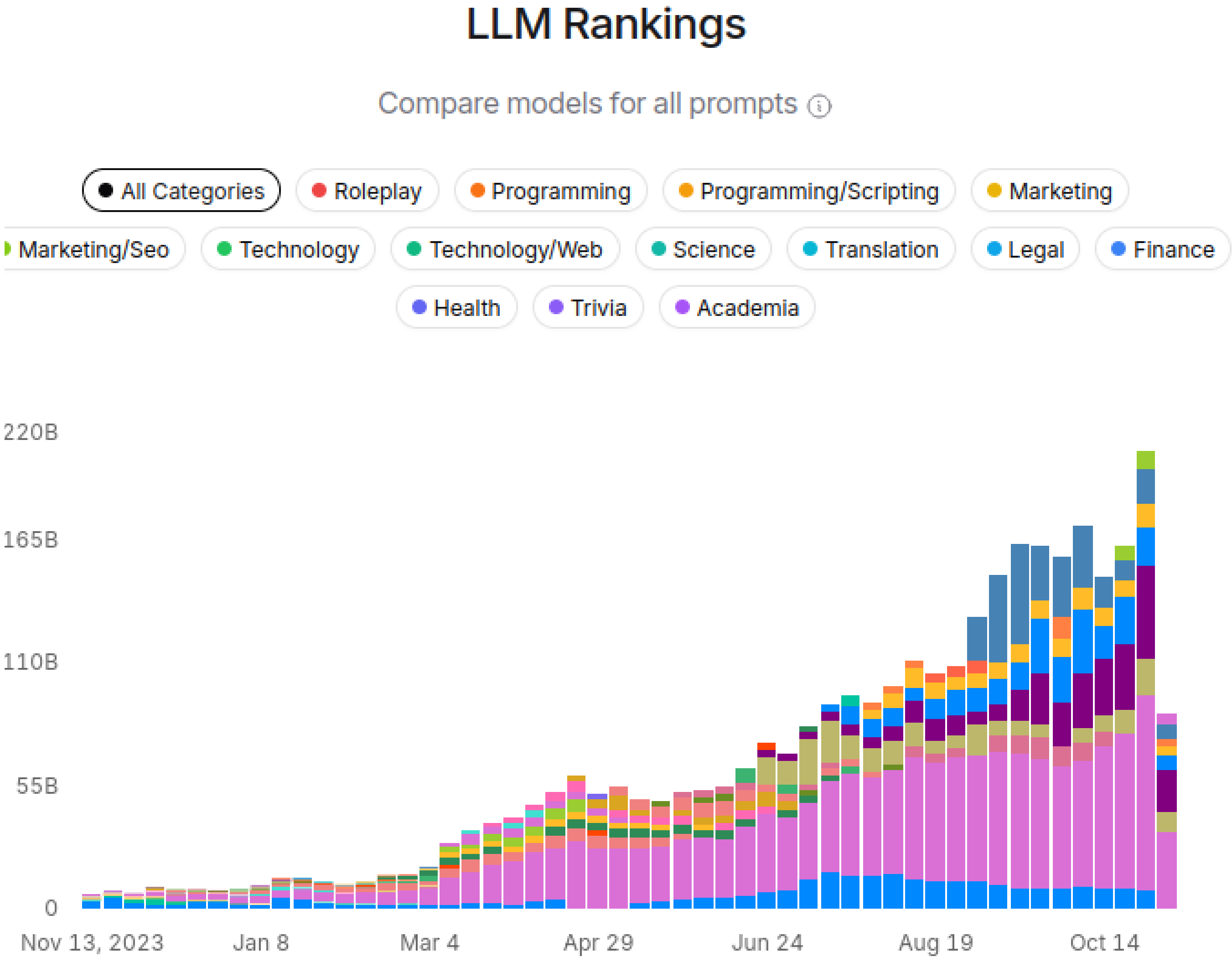
Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	1	ChatGPT-4o-latest (2024-09-03)	1340	+4/-3	33743	OpenAI	Proprietary	2023/10
1	1	o1-preview	1335	+4/-4	21071	OpenAI	Proprietary	2023/10
3	6	o1-mini	1308	+4/-4	23128	OpenAI	Proprietary	2023/10
3	4	Gemini-1.5-Pro-002	1303	+4/-4	15736	Google	Proprietary	Unknown
4	4	Gemini-1.5-Pro-Exp-0827	1299	+4/-3	32385	Google	Proprietary	2023/11
6	9	Grok-2-08-13	1290	+3/-3	40873	xAI	Proprietary	2024/3
6	3	Claude 3.5 Sonnet (20241022)	1286	+6/-6	7284	Anthropic	Proprietary	2024/4
6	11	Yi-Lightning	1285	+4/-4	20973	01 AI	Proprietary	Unknown
6	4	GPT-4o-2024-05-13	1285	+3/-3	102960	OpenAI	Proprietary	2023/10
10	15	GLM-4-Plus	1275	+4/-4	19922	Zhipu AI	Proprietary	Unknown
10	18	GPT-4o-mini-2024-07-18	1273	+4/-3	42661	OpenAI	Proprietary	2023/10
10	19	Gemini-1.5-Flash-002	1272	+5/-6	12379	Google	Proprietary	Unknown
10	26	Llama-3.1-Nemotron-70b-Instruct	1271	+5/-7	6228	Nvidia	Llama 3.1	2023/12
10	14	Gemini-1.5-Flash-Exp-0827	1269	+4/-4	25503	Google	Proprietary	2023/11

WHAT ARE PEOPLE ACTUALLY USING?

https://openrouter.ai/rankings



1.	Anthropic: Claude 3.5 Sonnet (self-moderated) > New Claude 3.5 Sonnet delivers better-than-Opus capabilities, fa...	44B tokens	↑14%
2.	Anthropic: Claude 3.5 Sonnet > New Claude 3.5 Sonnet delivers better-than-Opus capabilities, fa...	19.6B tokens	↑33%
3.	Google: Gemini Flash 1.5 > Gemini 1.5 Flash is a foundation model that performs well at a vari...	17.2B tokens	↑33%
4.	Google: Gemini 1.5 Flash-8B > Gemini 1.5 Flash-8B is optimized for speed and efficiency, offerin...	15B tokens	↑238%
5.	OpenAI: GPT-4o-mini > GPT-4o mini is OpenAI's newest model after [GPT-4 Omni](/mod...	14.9B tokens	↓15%
6.	Meta: Llama 3.1 70B Instruct > Meta's latest class of model (Llama 3.1) launched with a variety o...	10.8B tokens	↑21%
7.	Mistral: Mistral Nemo > A 12B parameter model with a 128k token context length built by ...	9.52B tokens	↑71%
8.	Meta: Llama 3.1 8B Instruct > Meta's latest class of model (Llama 3.1) launched with a variety o...	8.57B tokens	↑12%
9.	MythoMax 13B > One of the highest performing and most popular fine-tunes of Lla...	8.14B tokens	↓3%
10.	OpenAI: GPT-4o (2024-08-06) > The 2024-08-06 version of GPT-4o offers improved performanc...	4.62B tokens	↑6%



LLM Rankings

Compare models for all prompts ⓘ

AUDIO GENERATION (MUSIC)

Suno.ai

Udio.com

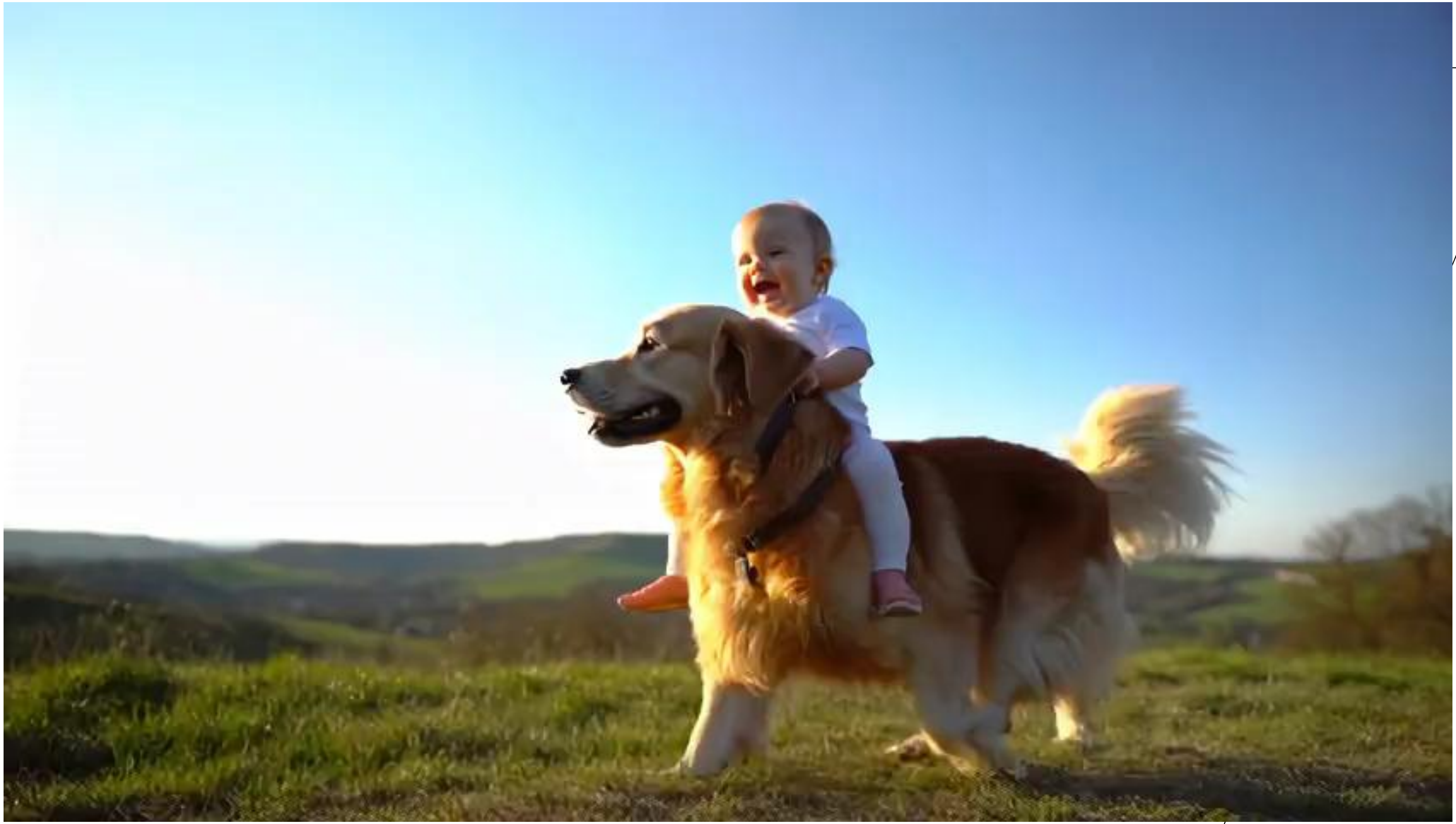
VIDEO GENERATION

- **Currently still expensive to generate videos**
- **Seems to be a competitive market, Chinese models seem to be leading in this space**

- **A lot of progress in the last months**

EXAMPLE 1 – META MOVIE GEN

EXAMPLE 2 - MOCHI



EXAMPLE 3 - MINIMAX

OTHER PROJECTS

- **moshi.chat (audio-to-audio)**
- **https://notebooklm.google/ (Google, generates podcasts)**
- **https://www.decart.ai/ Minecraft world simulator**

WHERE CAN I TRY ALL THESE THINGS?

- **locally (if you have the GPU and skills)**
- **Cloud rented GPU (e.g. vast.ai ~0.4€ per hour)**
- **Replicate (https://replicate.com, many models pay per use)**
- **OpenRouter (text-based-models, pay per use)**
- **Huggingface**
- **Various SaaS offerings**

HOW TO STAY INFORMED?

- **Huggingface Trending Models (https://huggingface.co/models)**

- **Leaderboards https://lmarena.ai/?leaderboard**
- **https://artificialanalysis.ai/ (also has leaderboards for image, video and speech generation)**

- **Reddit (/r/stablediffusion, /r/singularity, /r/ollama)**
- **X (https://x.com/emollick , the big studios (openai, metaai, anthropic and their employees), anonymous figures: @apples_jimmy, @kimmonismus, @tszzl)**

WORK TIME

Groups of 2-3. At least one laptop per group

Proposals:

1. **Create a little story with illustrations**
2. **Large Language Model LLM Comedy Arena (Input: topic to joke about, rate llms by how funny their jokes are)**
3. **Recipe bot (Input: photo of ingredients -> output recipe + photo of potential result)**
4. **Poker/Backgammon game recorder (Input: Video of poker game -> structured output of what happened)**

HOW TO WRITE CODE IF YOU DON'T KNOW HOW TO WRITE CODE

1. **Download and Install https://www.cursor.com/**
2. **Download Starting Point: https://github.com/JDemler/state-of-gen-ai/archive/refs/heads/main.zip**
3. **Select file you want to work in (for example comedy-club.html)**
4. **Press ctrl-i (opens cursor composer)**
5. **Tell composer what you want to have + click apply**
6. **Open file in browser (for example comedy-club.html)**
7. **Test the result, if it does not work as expected check Developer Tools for errors (F12 -> Console in Firefox, CTRL-SHIFT-J in Chrome)**
8. **Copy potential errors or your suggestions for improvements back into composer**
9. **Repeat until happy**

PRESENTATION OF RESULTS + DISCUSSION

THANK YOU

Jakob Demler

jdemler@curry-software.com

github.com/jdemler