

Digital humanities

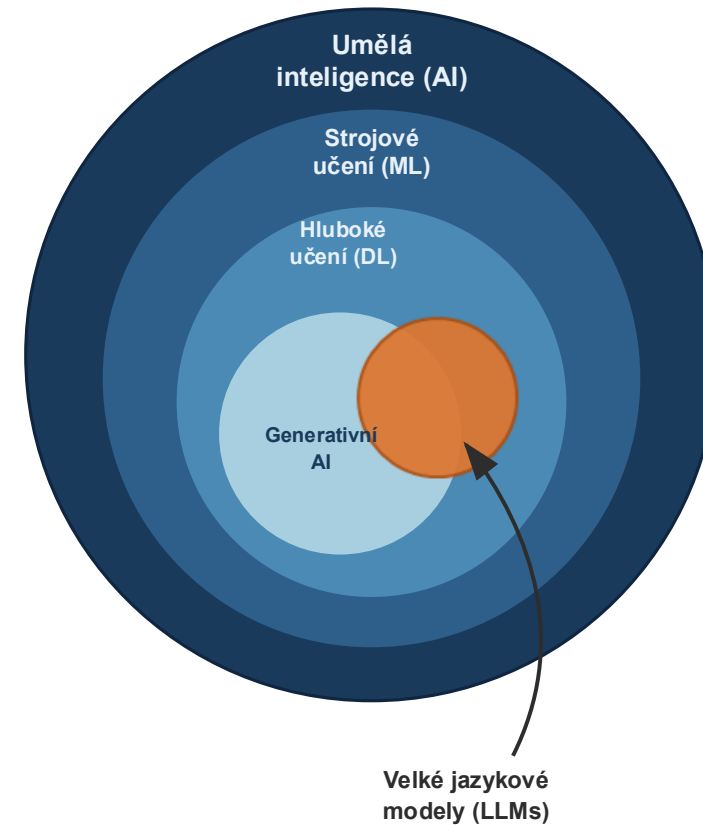
Tzv. umělá inteligence: teoretické základy a první kroky

Jindřich Marek

Tzv. umělá inteligence

základní pojmy (na základě četby)

- strojové učení (machine learning)
- hluboké učení (deep learning)
- neuronová síť (neural network)
- vnořování textu (text embedding)
- generativní umělá inteligence (generative artificial intelligence)



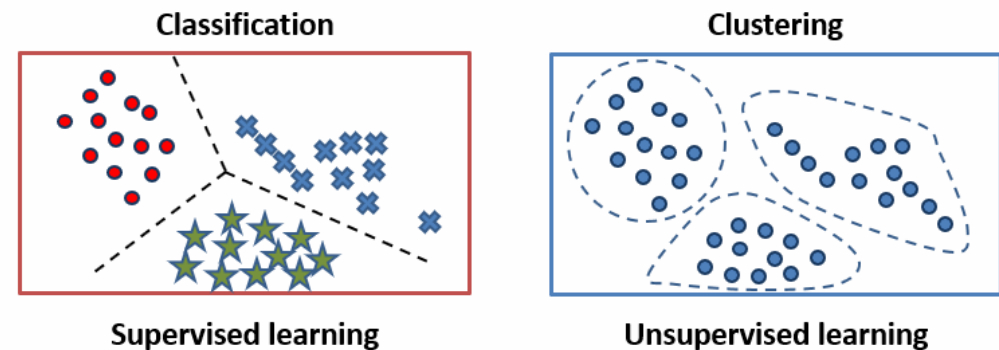
Co je generativní AI?

- schopnost vytvářet nový obsah – text, obrázky, kód

Jak se ale stroj naučil tvořit něco, co vypadá jako lidská myšlenka?

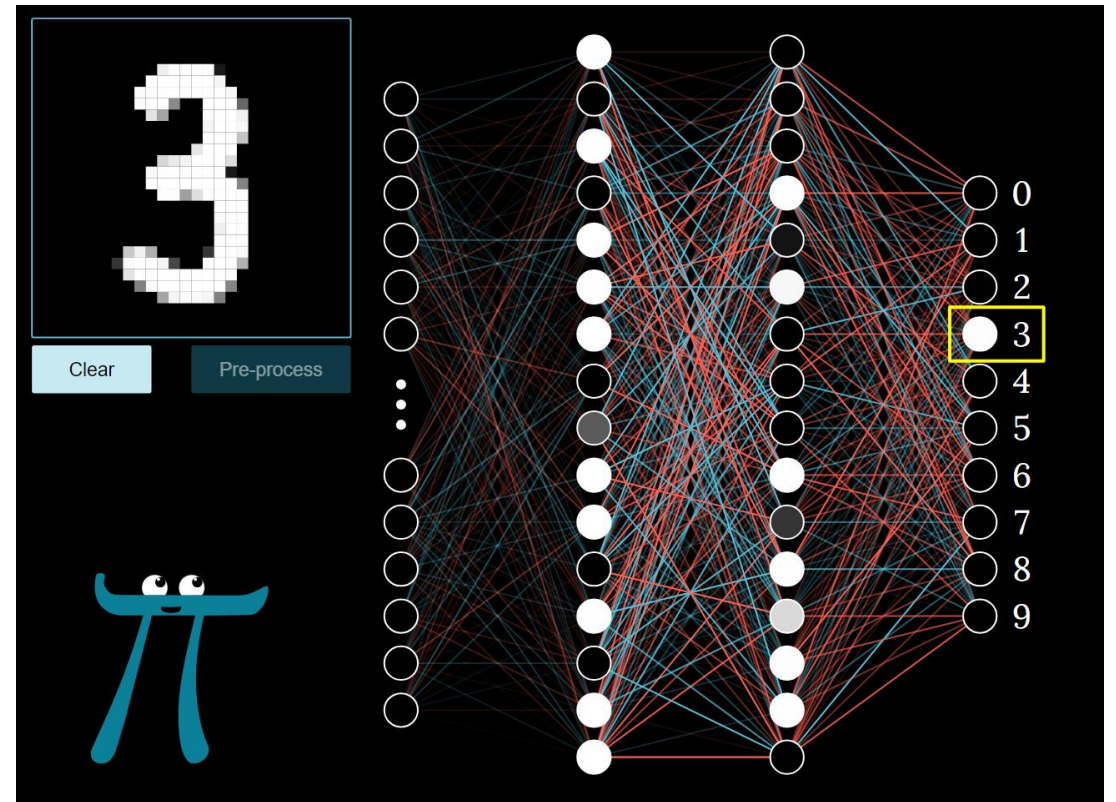
Princip: strojové učení

- učení z dat místo programování
- výstupy
 - klasifikace dat
 - vytváření předpovědí
- druhy
 - učení s učitelem (supervised learning)
 - učení bez učitele (unsupervised learning)
 - kombinace metod výše
 - zpětnovazebné učení (reinforcement learning)



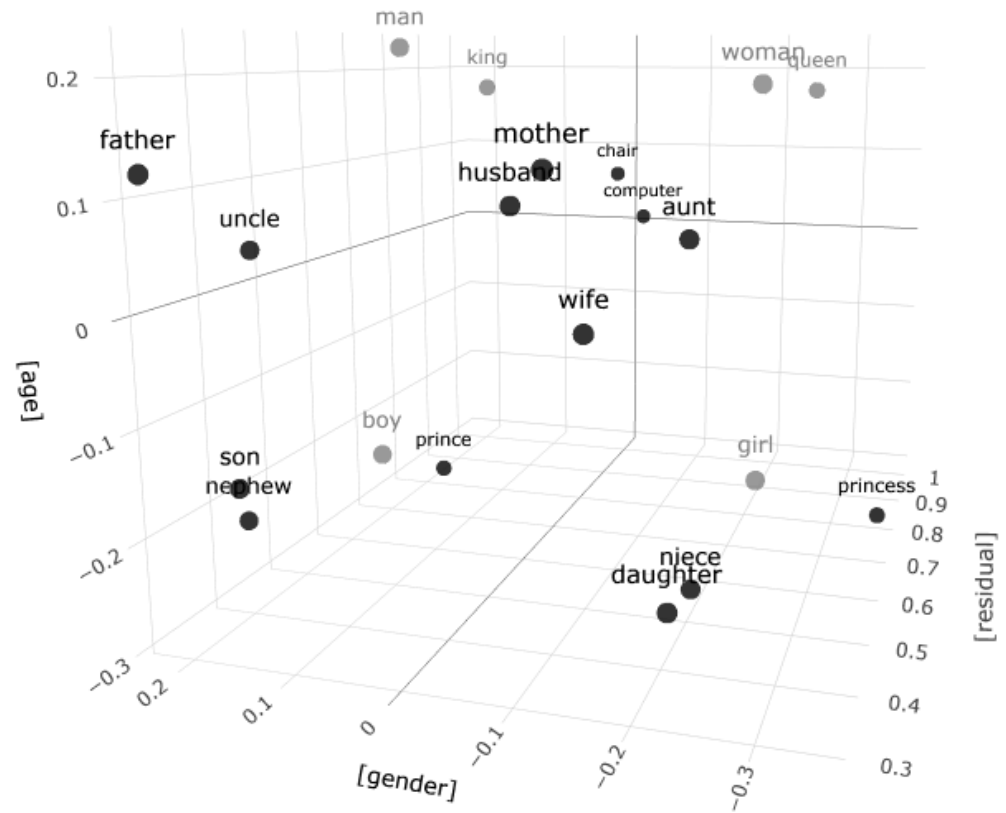
Základní stavební kámen: neuronové sítě

- struktura určená pro distribuované paralelní zpracování dat
- příklad
 - <https://www.3blue1brown.com/lessons/neural-network-analysis>



Překlad textu do čísel: vnořování textu

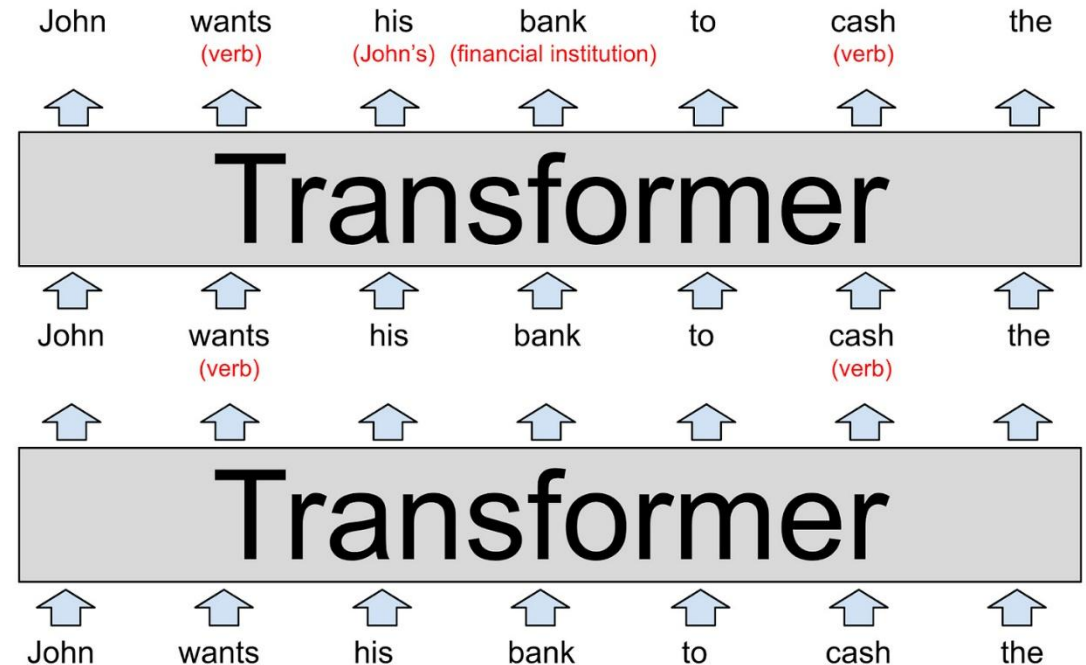
- převádí slova na body ve vícerozměrném prostoru (vektory)
 - slova s podobným významem jsou blízko u sebe
- <https://www.cs.cmu.edu/~dst/WordEmbeddingDemo/index.html>
- viz též <https://projector.tensorflow.org/>
 - zadat např. slovo intelligence



Architektura: transformer



- typ neuronové sítě, který přinesl mechanismus pozornosti (self-attention)
 - umožňuje při čtení slova zohlednit všechna ostatní slova ve větě najednou a posoudit, která z nich jsou pro kontext důležitá (to řeší různé významy slov)
 - umí zpracovávat text paralelně (rychle) a udržovat kontext i ve velmi dlouhých textech
 - přelomový článek od Google: *Attention Is All You Need* (2017)



Výsledek: velký jazykový model (LLM)

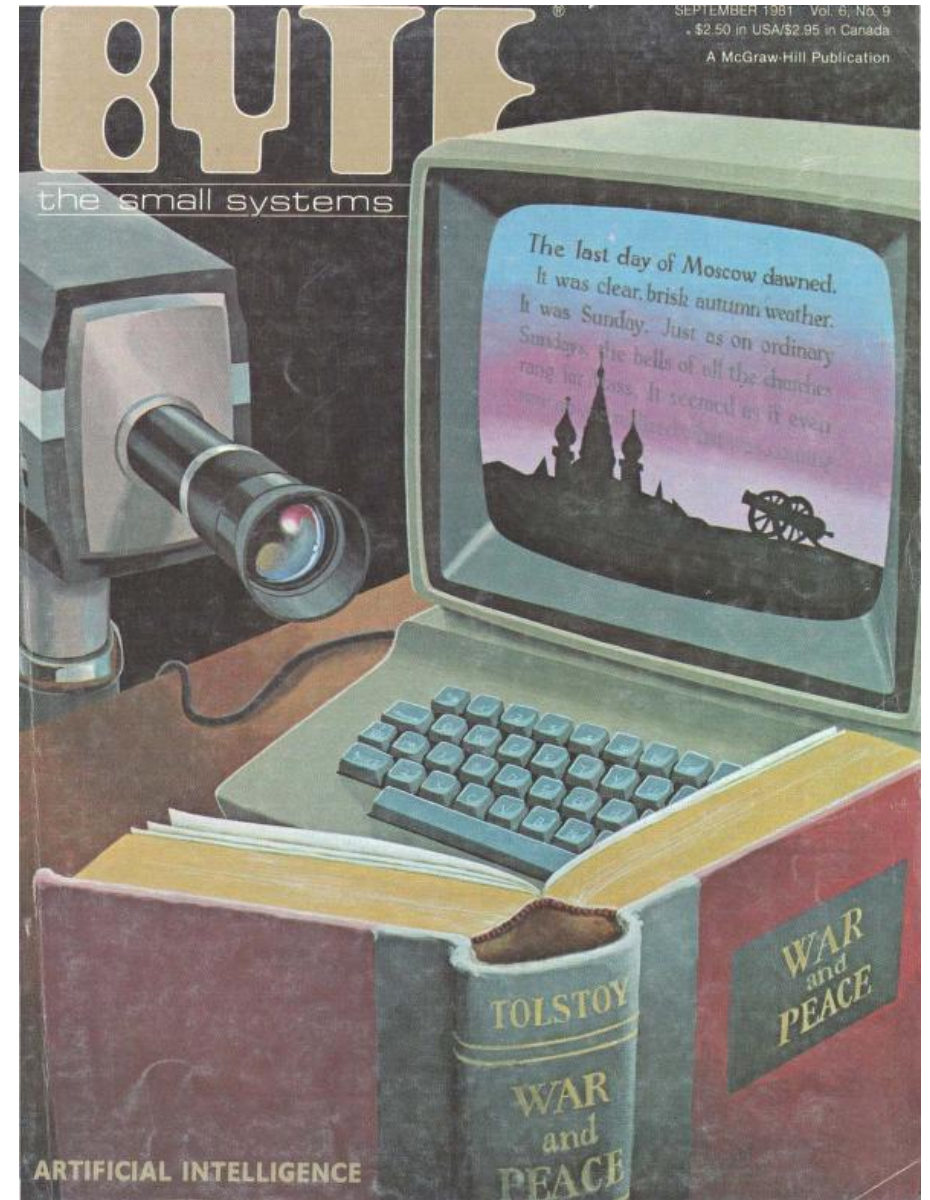
- vznik modelu:
 - s využitím architektury transformer vezmeme obrovské množství textu (typicky z internetu) naučíme ji předpovídat další slovo (pravděpodobnost)
- princip fungování:
 - vysoce pokročilé automatické doplňování textu na základě vnoření a pozornosti
 - čím je model větší (má větší počet parametrů), tím lépe „chápe“ logiku, faktografii a tón

Vytváření modelů

- datová sada (dataset)
 - data použita pro vytvoření modelu – většinou veřejně neodstupná
- váhy (weights)
 - vlastní model – komerční / s otevřenými vahami („open weights“) = váhy dostupné ke stažení

Vlastnosti modelů

- multimodalita
 - na vstupu či výstupu nejen text, ale i obraz, audio
- jednoduché / směs expertů (dense / mixture of experts)
- schopnost řešit složité úlohy (simulované vyvozování)
 - „myšlení“ před definitivním výsledkem
- volání funkcí (využití nástrojů)
 - možnost využít nástroj, např. Python
- online / lokální
 - webová služba, nebo využití lokálních vah
- počet parametrů
 - v miliardách (B), vliv na využití paměti (VRAM, příp. RAM)
 - kvantizace: snížení velikosti + snaha o zachování kvality



Příklady modelů (duben 2026)

komerční modely online:

- OpenAI
 - GPT-5.4 Chat
- Anthropic
 - Claude Sonnet 4.6
 - Claude Opus 4.6
- Google
 - Gemini 3.1 Flash
 - Gemini 3.1 Pro
- Cohere
 - Command A

volně dostupné modely (zde ty menší):

- Google
 - Gemma 3 (270M, 1B, 4B, 12B, 27B)
 - Gemma 4 (E2B, E4B, 26B-A4B, 31B)
- Alibaba Cloud
 - Qwen 3.5 (0.8B, 2B, 4B, 9B, 27B, 35B-A3B)
- MistralAI
 - Mistral Small 3.2 (24B)
 - Ministral 3 (3B, 8B, 14B)
- Cohere
 - Aya Expanse (8B, 32B)

OpenRouter: komerční modely (přístup)

- <https://openrouter.ai>

Input Modalities

- Text
- Image
- File
- Audio
- Video
- New

Output Modalities

- Text
- Image
- Audio
- Embeddings
- New

Context length

Prompt pricing

Series

Categories

Supported Parameters

Distillable

Zero Data Retention

Models

642 models

Most Popular

Model Name	Weekly Tokens	Input (\$/1M)	Output (\$/1M)	Context	Released
 MiniMax: MiniMax M2.5	1.91T	\$0.295	\$1.20	196 608	Feb 12, 2026
 Google: Gemini 3 Flash Preview	1.09T	\$0.50	\$3	1 048 576	Dec 17, 2025
 DeepSeek: DeepSeek V3.2	866B	\$0.25	\$0.40	163 840	Dec 1, 2025
 MoonshotAI: Kimi K2.5	825B	\$0.45	\$2.20	262 144	Jan 27, 2026
 Anthropic: Claude Opus 4.6	802B	\$5	\$25	1 000 000	Feb 4, 2026
 StepFun: Step 3.5 Flash (free) 	730B	\$0	\$0	256 000	Jan 30, 2026
 Anthropic: Claude Sonnet 4.6	701B	\$3	\$15	1 000 000	Feb 17, 2026
 xAI: Grok 4.1 Fast	631B	\$0.20	\$0.50	2 000 000	Nov 19, 2025
 Arcee AI: Trinity Large Preview (free) 	569B	\$0	\$0	131 000	Jan 27, 2026
 Google: Gemini 2.5 Flash	541B	\$0.30	\$2.50	1 048 576	Jun 17, 2025
 Anthropic: Claude Sonnet 4.5	531B	\$3	\$15	1 000 000	Sep 29, 2025
 OpenAI: gpt-oss-120b	484B	\$0.039	\$0.19	131 072	Aug 5, 2025
 Google: Gemini 2.5 Flash Lite	425B	\$0.10	\$0.40	1 048 576	Jul 22, 2025
 Z.ai: GLM 5	335B	\$0.80	\$2.56	202 752	Feb 11, 2026

Hugging Face: zdroj lokálních modelů

- <https://huggingface.co>



Main Tasks Libraries Languages Licenses Other

Tasks

Text Generation

Any-to-Any

Image-Text-to-Text

Image-to-Text

Image-to-Image

Text-to-Image

Text-to-Video

Text-to-Speech

+ 44

Parameters



Libraries

PyTorch

TensorFlow

JAX

Transformers

Diffusers

sentence-transformers

Safetensors

ONNX

GGUF

Transformers.js

MLX

MLX

+ 41

Apps

vLLM

llama.cpp

MLX LM

LM Studio

Ollama

Jan

Draw Things

+ 8

Inference Providers

Groq

Novita

Cerebras

SambaNova

Nscale

fal

Hyperbolic

Together AI

+ 10

Models 727,363

Filter by name

Full-text search

Inference Available

Sort: Trending

Qwen/Qwen3.5-9B

Image-Text-to-Text · 10B · Updated 6 days ago · 868k · 583

Qwen/Qwen3.5-4B

Image-Text-to-Text · 5B · Updated 6 days ago · 349k · 290

HauhausCS/Qwen3.5-9B-Uncensored-HauhausCS-Aggressive

9B · Updated 4 days ago · 75.5k · 196

Qwen/Qwen3.5-27B

Image-Text-to-Text · 28B · Updated 11 days ago · 698k · 608

unsloth/Qwen3.5-4B-GGUF

Image-Text-to-Text · 4B · Updated 6 days ago · 206k · 113

Jackrong/Qwen3.5-27B-Claude-4.6-Opus-Reasoning-Dist...

Text Generation · 27B · Updated about 11 hours ago · 36.4k · 113

crownelius/Crow-9B-Opus-4.6-Distill-Heretic_Qwen3.5

9B · Updated 1 day ago · 26.2k · 95

unsloth/Qwen3.5-0.8B-GGUF

Image-Text-to-Text · 0.8B · Updated 6 days ago · 122k · 76

unsloth/LTX-2.3-GGUF

Image-to-Video · 21B · Updated about 18 hours ago · 24.8k · 65

janhq/Jan-code-4b

Text Generation · 4B · Updated 4 days ago · 1.29k · 62

Qwen/Qwen3.5-0.8B

Image-Text-to-Text · 0.9B · Updated 6 days ago · 406k · 317

unsloth/Qwen3.5-9B-GGUF

Image-Text-to-Text · 9B · Updated 6 days ago · 505k · 248

Jackrong/Qwen3.5-27B-Claude-4.6-Opus-Reasoning-Dist...

Text Generation · 28B · Updated about 11 hours ago · 9.21k · 215

Qwen/Qwen3.5-2B

Image-Text-to-Text · 2B · Updated 6 days ago · 227k · 152

microsoft/Phi-4-reasoning-vision-15B

Image-Text-to-Text · 15B · Updated 4 days ago · 14.3k · 108

unsloth/Qwen3.5-27B-GGUF

Image-Text-to-Text · 27B · Updated 3 days ago · 477k · 263

FireRedTeam/FireRed-OCR

Image-to-Text · 2B · Updated 4 days ago · 5.23k · 105

perplexity-ai/pplx-embed-v1-0.6b

Feature Extraction · 0.6B · Updated 5 days ago · 155k · 147

HauhausCS/Qwen3.5-4B-Uncensored-HauhausCS-Aggressive

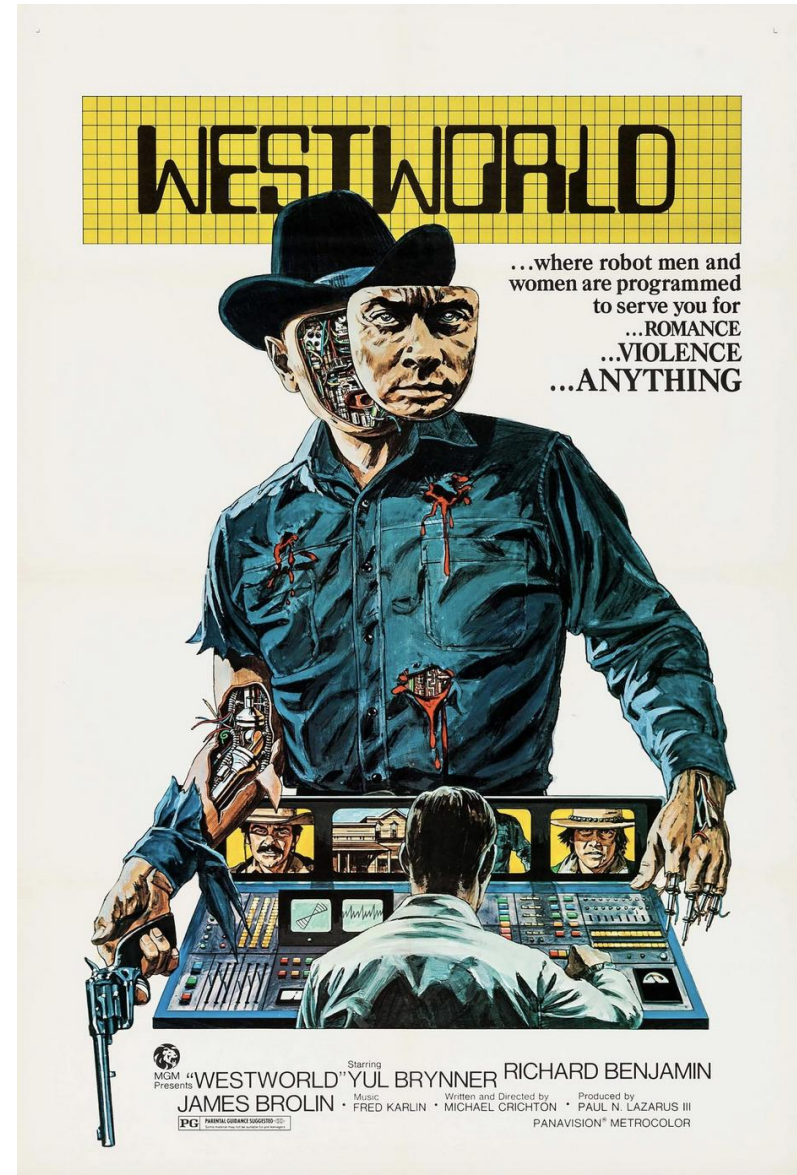
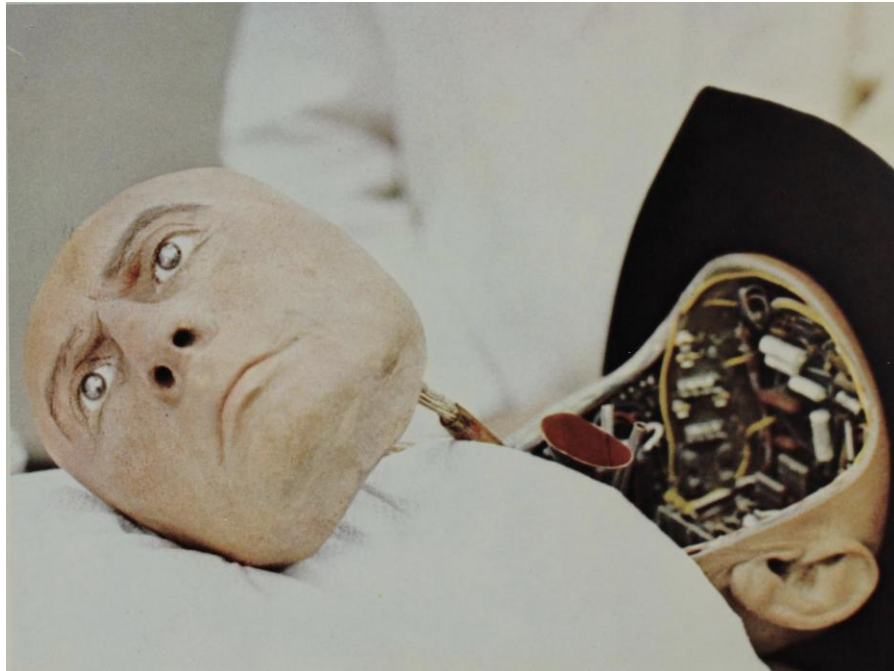
4B · Updated 4 days ago · 34.4k · 64

TeichAI/GLM-4.7-Flash-Claude-Opus-4.5-High-Reasonin...

30B · Updated 14 days ago · 98.7k · 443

Nástroje pro práci s LLM

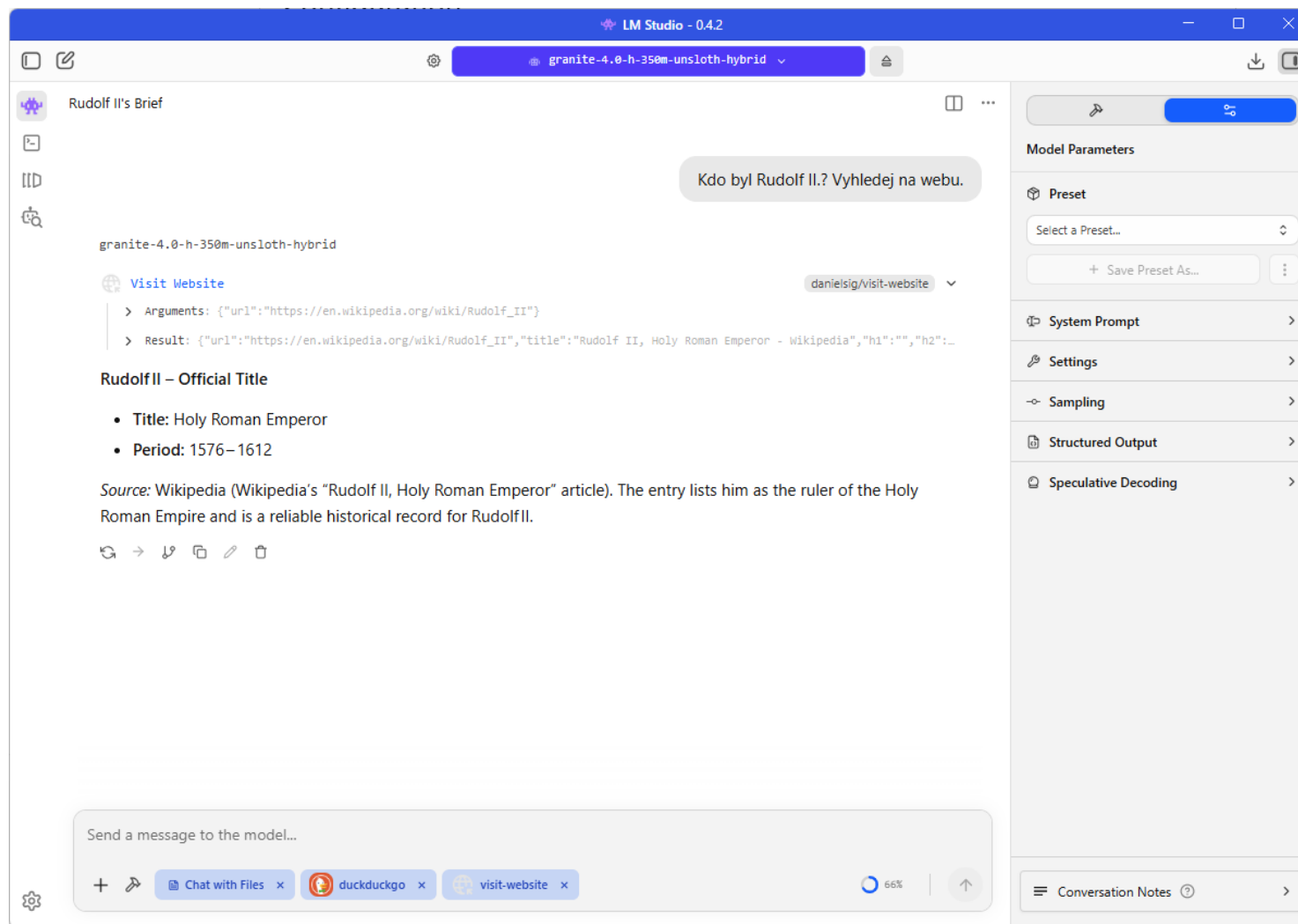
- příklady nástrojů
- základy práce s API



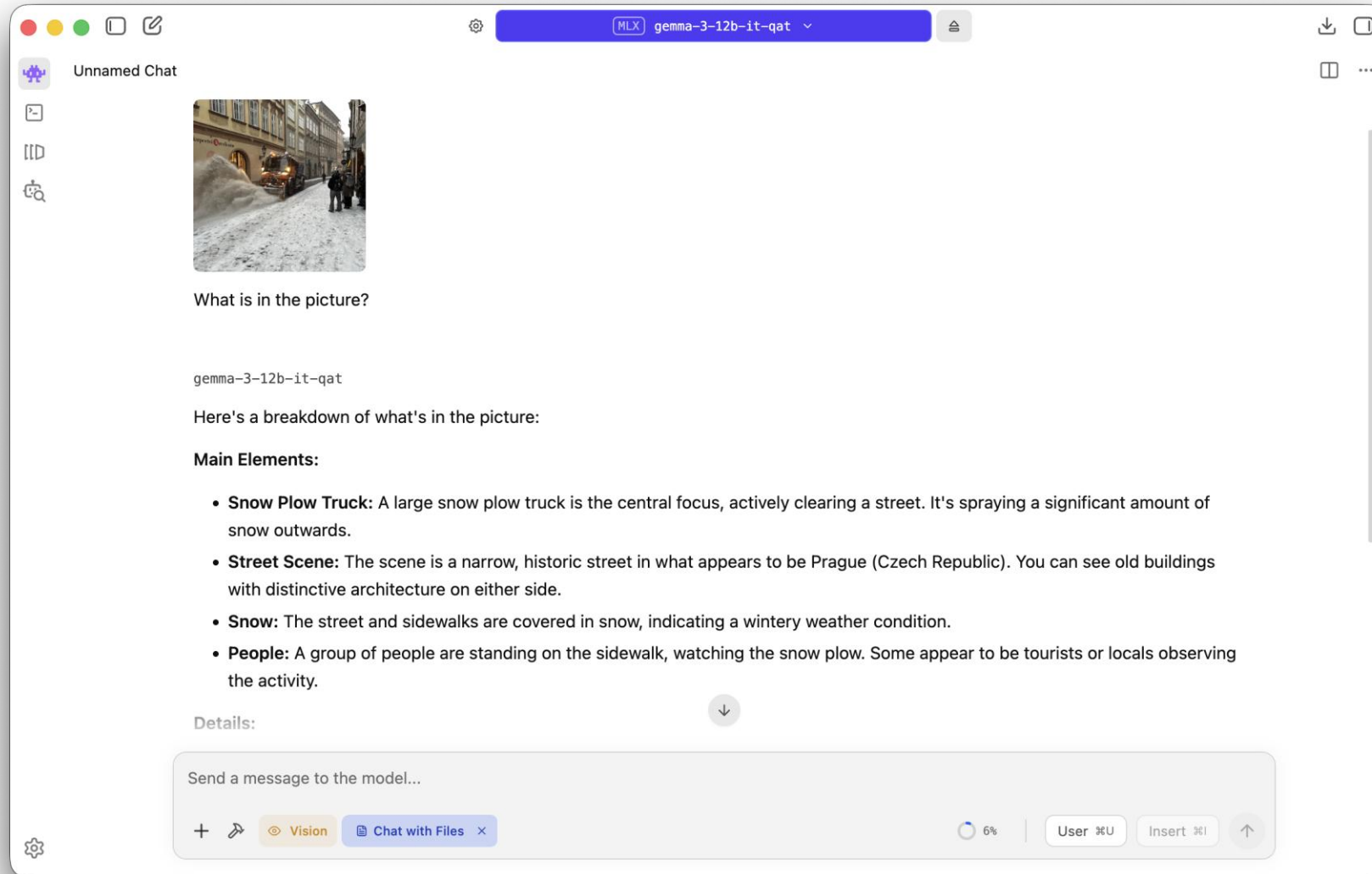
Příklady nástrojů

- LM Studio
 - aplikace pro spouštění lokálních modelů (grafické prostředí)
- Python
 - programovací jazyk pro skripty a integraci

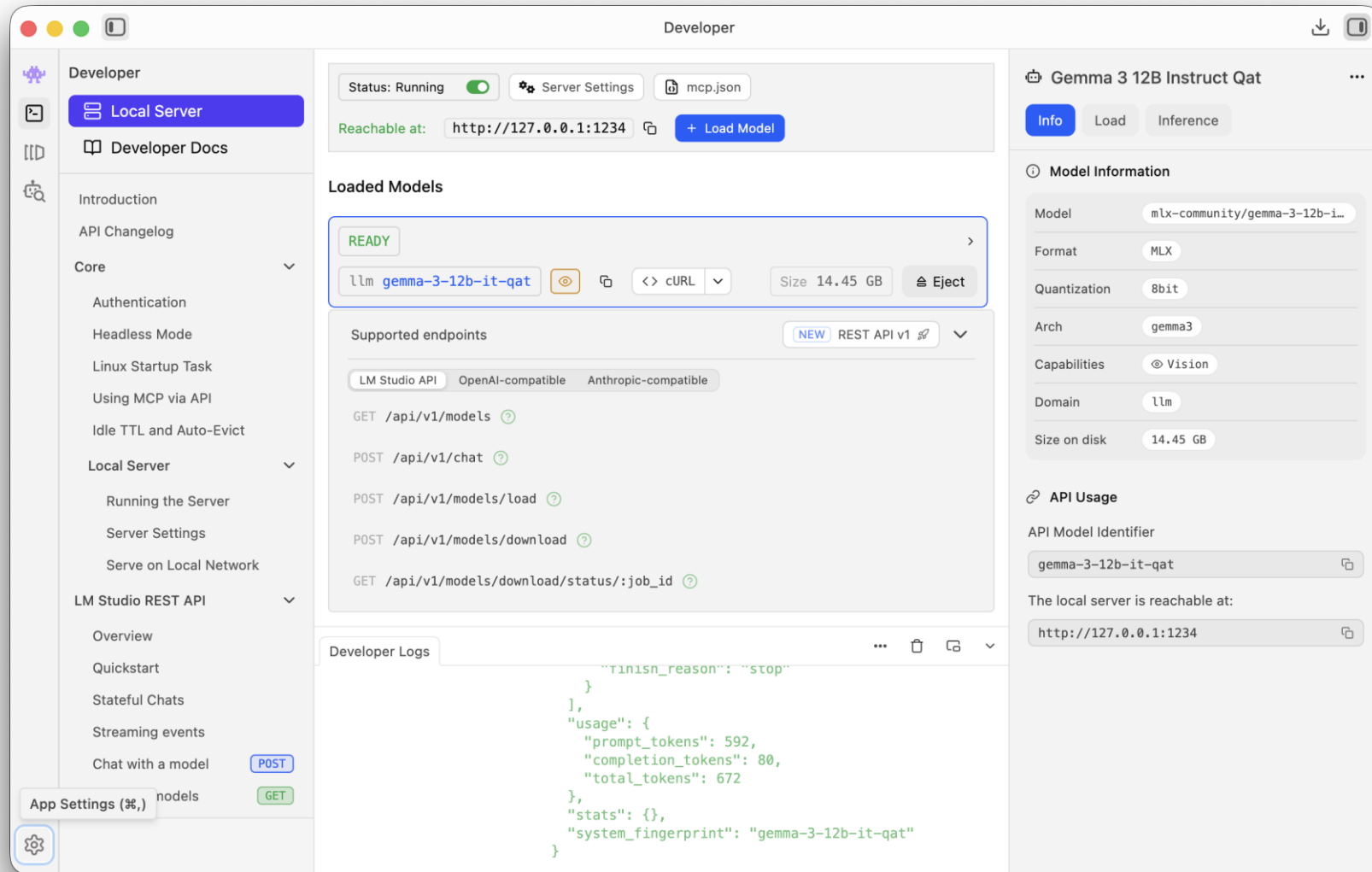
LM Studio: lokální model s vyhledáváním



LM Studio: dotazování k obrázkům



LM Studio: lokální server poskytující API



Práce s API v Pythonu

Generate text from a simple prompt

python ↕ 

```
1 from openai import OpenAI
2 client = OpenAI()
3
4 response = client.responses.create(
5     model="gpt-5-nano",
6     input="Write a one-sentence bedtime story about a unicorn."
7 )
8
9 print(response.output_text)
```

API, nebo webové rozhraní?

- API
 - automatizace, zpracování velkého množství dat, integrace do workflow
 - možnost „zkrocení“ modelů: struktura výstupu, validace
- webové rozhraní
 - rychlé vyzkoušení, jednotlivé dotazy, pro méně technické uživatele