

Digital humanities

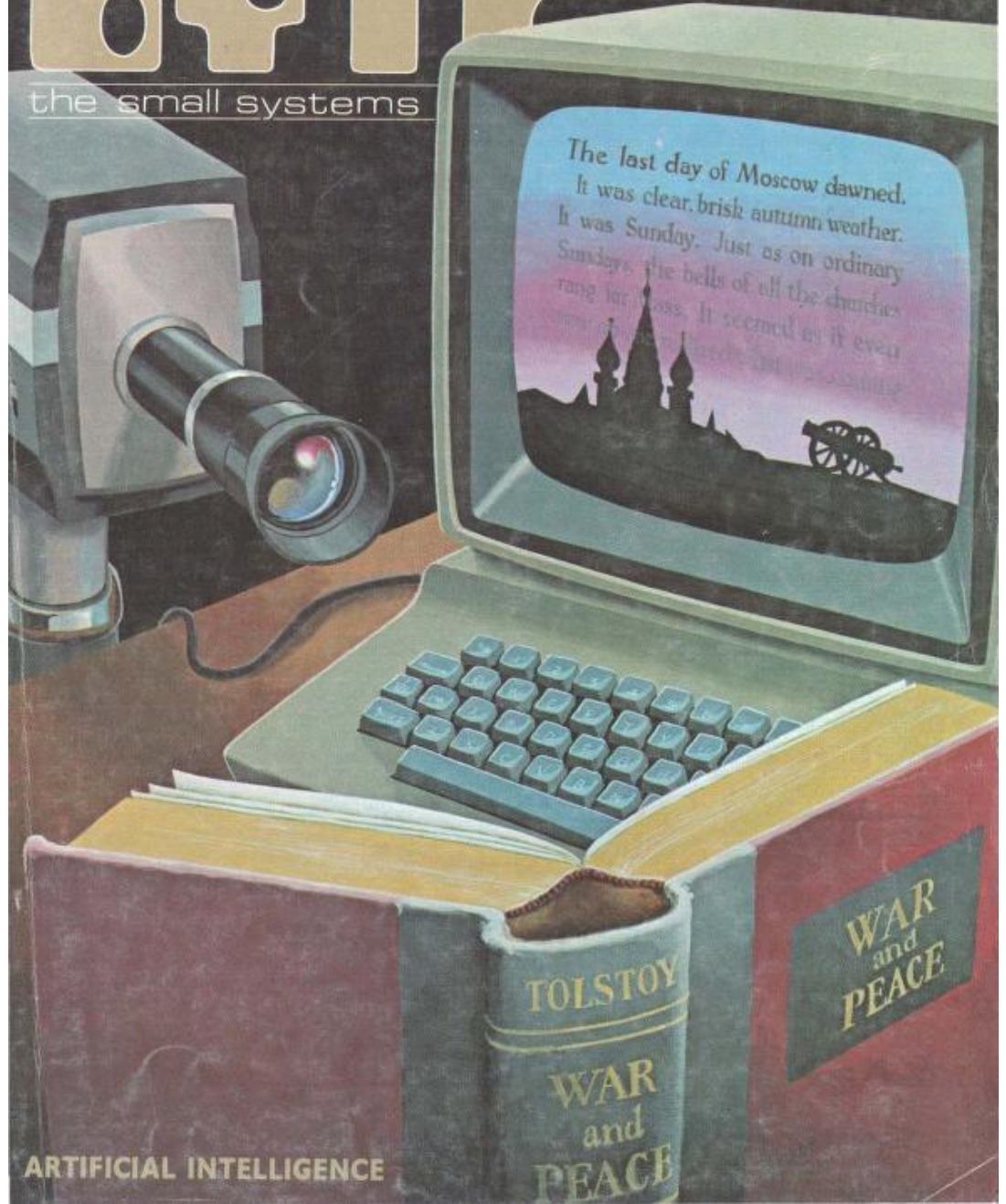
Velké jazykové modely a práce s textem

Jindřich Marek

BYTE

the small systems

SEPTEMBER 1981 Vol. 6, No. 9
\$2.50 in USA/\$2.95 in Canada
A McGraw-Hill Publication



ARTIFICIAL INTELLIGENCE

Pojmy (na základě četby)

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

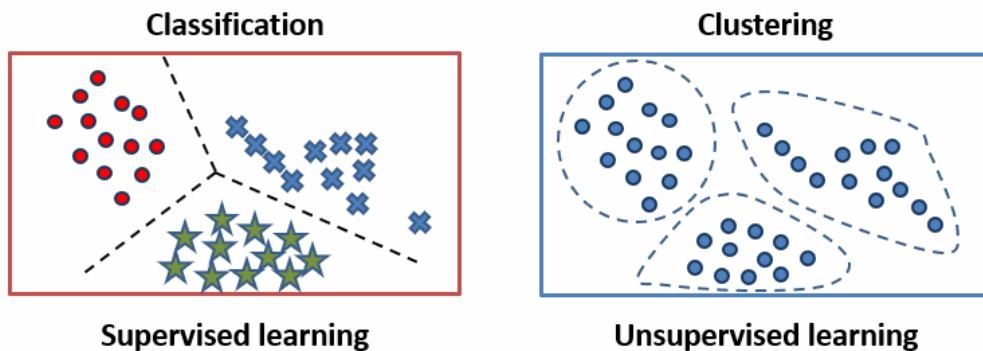
Strojové učení

- druhy

- učení s učitelem (supervised learning)
- učení bez učitele (unsupervised learning)
- kombinace metod výše
- zpětnovazebné učení (reinforcement learning)

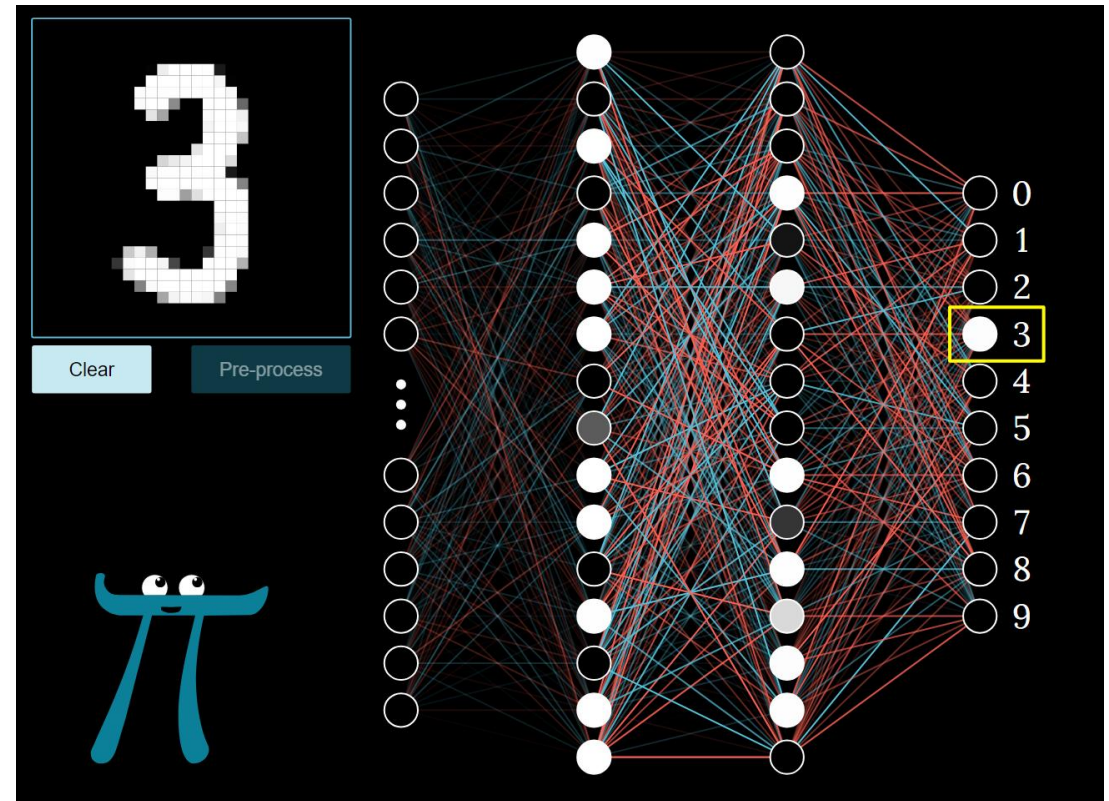
- výstupy

- klasifikace dat
- vytváření předpovědí



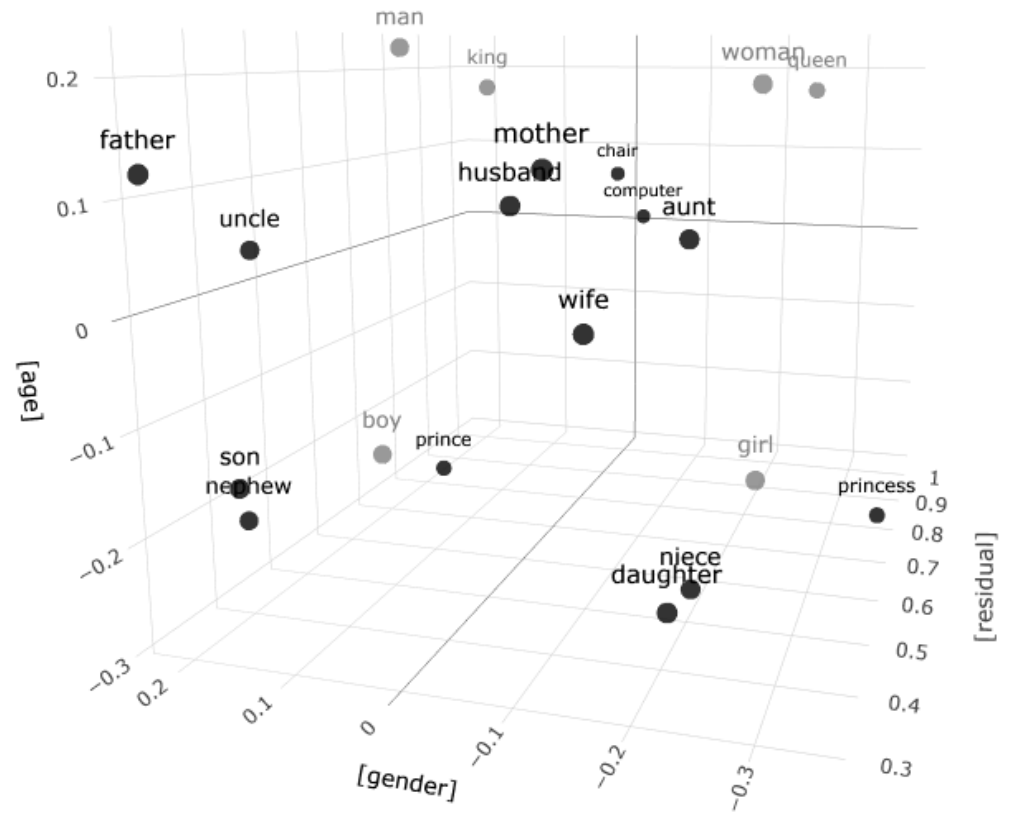
Neuronové sítě

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

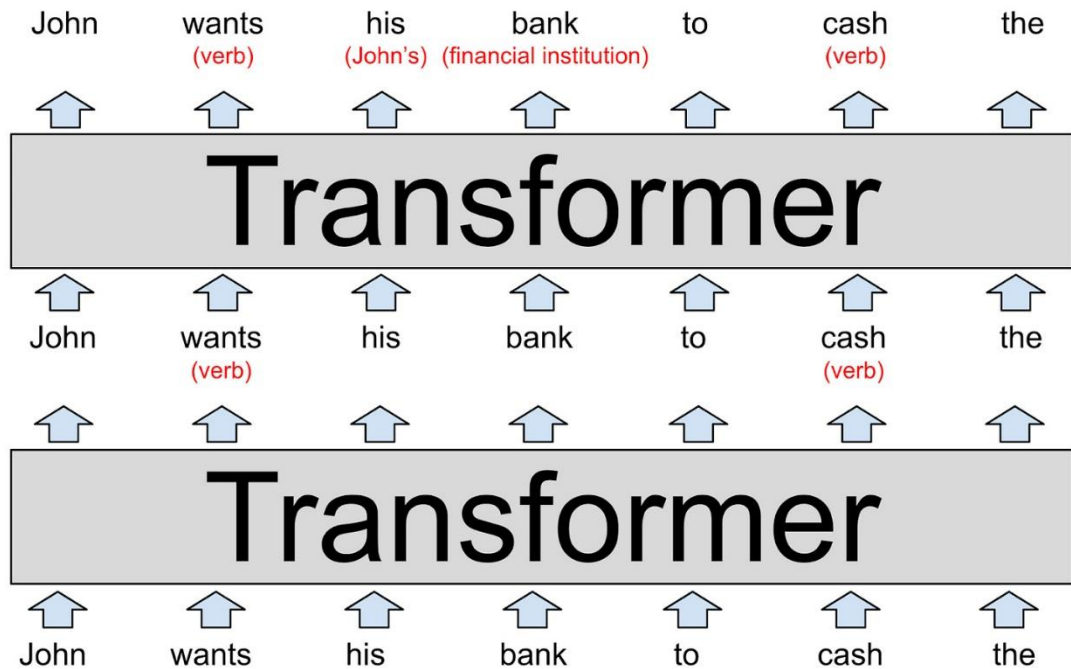


Vnořování textu (text embedding)

- <https://www.cs.cmu.edu/~dst/WordEmbeddingDemo/index.html>
 - viz též <https://projector.tensorflow.org/>



Transformers



Podstata velkých jazykových modelů

- kódování textu v mnoharozměrném prostoru
- transformers
- kvantizace
- využití paměti (VRAM vs. RAM, offloading)
 - Llama 3 8B ve formátu GGUF s kvantizací Q8_0 spotřebovává průměrně 8-10 GB (V)RAM
 - rychlost generování na relativně rychlém procesoru (Ryzen 5) je při nepřiliš ztrátových kvantizacích obvykle cca 2-3,5 tokenů za sekundu

Vytváření modelů

- trénování modelů: vznik natrénovaného modelu (pre-trained)
 - dataset
 - weights

Vylepšování modelů

- ladění modelů: úprava modelu pro konkrétní úkol
 - dataset
 - weights: LoRA (Low-Rank Adaptation)
- je třeba si dát pozor na tzv. overfitting

Příklady velkých jazykových modelů

- online
 - OpenAI GPT
 - GPT-4 (Turbo)
 - GPT-3.5 (Turbo)
 - Anthropic Claude 3
 - Opus
 - Sonnet
 - Haiku
 - Google Gemini 1.5 Pro
 - ...
- lokální modely (open weights)
 - Llama 2
 - 7B, 13B, 70B
 - Llama 3
 - 8B, 70B
 - Mistral 7B
 - Mixtral 8x7B
 - Mixtral 8x22B
 - ...



How can I help you today?

Plan an itinerary

to experience the wildlife in the Australian outback

Explain this code:

```
"cat config.yaml | awk NF"
```

Make up a story

about Sharky, a tooth-brushing shark superhero

Design a programming game

teach basics in a fun way

Message ChatGPT...





You

Who was Rudolf II?



ChatGPT

Rudolf II was a Holy Roman Emperor who ruled from 1576 until his abdication in 1611. He was born on July 18, 1552, in Vienna, Austria, and was a member of the Habsburg dynasty. Rudolf II is often remembered for his patronage of the arts and sciences, particularly during the Renaissance period. He was known for his interest in alchemy, astrology, and the occult, and he gathered a vast collection of artworks and curiosities in Prague, where he moved his court in 1583. However, his reign was also marked by political instability and conflicts within the Holy Roman Empire, including tensions with Protestant princes and the outbreak of the Thirty Years' War. Rudolf II died on January 20, 1612, in Prague, Czech Republic.

Message ChatGPT...



huggingface.co: zdroj lokálních modelů

Tasks Libraries Datasets Languages Licenses
Other

Multimodal

 Image-Text-to-Text

 Visual Question Answering

 Document Question Answering

Computer Vision

 Depth Estimation  Image Classification

 Object Detection  Image Segmentation

 Text-to-Image  Image-to-Text

 Image-to-Image  Image-to-Video

 Unconditional Image Generation

 Video Classification  Text-to-Video


Models 604,133

[new](#) Full-text search

↑↓ Sort: Trending

 **mistral-community/Mixtral-8x22B-v0.1**

 Text Generation • Updated 5 days ago • ↓ 28.7k • ♥ 538

 **CohereForAI/c4ai-command-r-plus**


 Text Generation • Updated 6 days ago • ↓ 150k • ♥ 1.11k

 **openbmb/MiniCPM-V-2**


 Visual Question Answering • Updated about 13 hours ago • ↓ 1.69k • ♥ 259

 **NexaAIDev/Octopus-v2**

 Text Generation • Updated 4 days ago • ↓ 10.2k • ♥ 714

 **HuggingFaceH4/zephyr-orpo-141b-A35b-v0.1**


 Text Generation • Updated 4 days ago • ↓ 3.11k • ♥ 174

 **parler-tts/parler_tts_mini_v0.1**

Prostředí pro lokální modely

- online rozhraní
 - <https://labs.perplexity.ai>
 - <https://sdk.vercel.ai>
 - ...
- online notebooky
 - Google Colab
 - Kaggle
 - ...
- lokální
 - LM Studio
 - program
 - Text generation web UI
 - běží na lokálním PC v prohlížeči
 - ollama
 - ...
 - Pro provoz modelů se 7B parametrů byste měli mít k dispozici alespoň 8 GB paměti RAM, pro provoz modelů s 13B parametrů 16 GB a pro provoz modelů s 33B parametrů 32 GB.

LLM served by Perplexity Labs
Hello! How can I help you?

 Copy



0.00 SEC

mixtral-8x7b-instruct



Ask anything...





AI

Chat

Prompt

Docs

Feedback

Login with Vercel

+ New Chat

Sign in to save your playgrounds

OpenAI gpt-3.5-turbo



Meta llama-v2-70b-chat



OpenAI / gpt-3.5-turbo

OpenAI's most capable and cost effective model in the GPT-3.5 family optimized for chat purposes, but also works well for traditional completions tasks.

Context	4 096 tokens
Input Pricing	\$1.50 / million tokens
Output Pricing	\$2.00 / million tokens

[Model Page](#) [Pricing](#) [Website](#)

Meta / llama-v2-70b-chat

70 billion parameter open source model by Meta fine-tuned for chat purposes served by Fireworks. LLaMA v2 was trained on more data (~2 trillion tokens) compared to LLaMA v1 and supports context windows up to 4k tokens.

Context	4 096 tokens
Input Pricing	\$0.70 / million tokens
Output Pricing	\$2.80 / million tokens

[Model Page](#) [Pricing](#) [Website](#)

Send a message

Sync Chats

Send a message

Sync Chats

[AI Policy](#)
[Privacy Policy](#)

Search for models by keyword or paste any HuggingFace repo URL .

Supports any

Llama

Mistral

Phi-2

Falcon

StarCoder

StableLM

GPT-NeoX

gguf ⓘ model file on Hugging Face

Welcome to LM Studio!

[Release Notes \(v0.2.19\)](#)

LM Studio enables you develop and experiment with Large Language Models (LLMs) in your local computer environment, fully offline.

Tip: Start with very small LLMs and move up to larger models depending on your hardware's capabilities.

Search Search and download compatible model files

AI Chat Chat with local LLMs fully offline

Multi Model Load and prompt multiple local LLMs simultaneously

Local Server Run an OpenAI-like HTTP server on localhost

My Models Manage your downloaded models

• Join [LM Studio's Discord Server](#) to discuss models, prompts, workflows and more.

Stability AI

3B

StableLM

Requires 8GB+ RAM

Stable Code Instruct 3B

Stable Code Instruct 3B is a decoder-only language model with 2.7 billion parameters, developed from the stable-code-3b. It has been trained on a combination of publicly available and synthetic datasets, with the latter generated through... [Show all](#)

File Size 2.97 GB

Less Compressed

Q8_0

↓ Download

Published by Bartowski on Hugging Face

Local Inference Server

Start a local HTTP server that mimics select OpenAI API endpoints.

Supported endpoints: /v1/chat/completions, /v1/embeddings, /v1/models

API Documentation: <https://lmstudio.ai/docs/local-server> (open in browser)

Configuration

Server Port: 1234

Cross-Origin-Resource-Sharing (CORS): ON OFF

Request Queuing: ON OFF

Verbose Server Logs: ON OFF

Start Server Stop Server

Embedding Model Settings

Learn More About Embeddings

Load a text embedding model and utilize it through the POST /v1/embeddings endpoint.

ChristianAzinn • labse 109M Q8_0 gguf

Eject

Offload entire model to VRAM

LM Studio will set the context length and batch size to 512. Modifying this will be supported in future versions.

Examples

hello world (curl) chat (python) ai assistant (python) vision (python) embeddings (python)

```
# Make sure to `pip install openai` first
from openai import OpenAI
client = OpenAI(base_url="http://localhost:1234/v1", api_key="lm-studio")

def get_embedding(text, model="ChristianAzinn/labse-gguf"):
    text = text.replace("\n", " ")
    return client.embeddings.create(input = [text], model=model).data[0].embedding
```

Copy Code

Server logs Server running on port 1234 (logs are saved into /tmp/lmstudio-server-log.txt)

Filter logs...

Open Logs

Clear (Ctrl+K)

```
[2024-04-16 18:47:28.373] [INFO] [LM STUDIO SERVER] Verbose server logs are ENABLED
[2024-04-16 18:47:28.387] [INFO] [LM STUDIO SERVER] Success! HTTP server listening on port 1234
[2024-04-16 18:47:28.389] [INFO] [LM STUDIO SERVER] Supported endpoints:
[2024-04-16 18:47:28.390] [INFO] [LM STUDIO SERVER] -> GET http://localhost:1234/v1/models
[2024-04-16 18:47:28.390] [INFO] [LM STUDIO SERVER] -> POST http://localhost:1234/v1/chat/completions
[2024-04-16 18:47:28.390] [INFO] [LM STUDIO SERVER] -> POST http://localhost:1234/v1/completions
[2024-04-16 18:47:28.391] [INFO] [LM STUDIO SERVER] -> POST http://localhost:1234/v1/embeddings <----- NEW!
[2024-04-16 18:47:28.392] [INFO] [LM STUDIO SERVER] Model loaded: TheBloke/Mistral-7B-Instruct-v0.2-GGUF/mistral-7b-instruct-v0.2.Q4_K_M.gguf
[2024-04-16 18:47:28.392] [INFO] [LM STUDIO SERVER] Logs are saved into C:\tmp\lmstudio-server-log.txt
```

Server Model Settings

Reset to Default Settings

Preset

Mistral Instruct

Danger Zone

Discard Changes

Override Preset

System Prompt

20 tokens

Below is an instruction that describes a task. Write a response that appropriately completes the request.

Ask before overriding

Advanced Configuration

Config for Server Configuration

New Preset

Export to JSON

Hardware Settings

GPU Acceleration

Show Help

GPU Offload

n_gpu_layers

0



Quick GPU Offload Settings

Tools

Model Inspector

Context Overflow Policy

Zobrazit plochu

Chat

Default

Notebook

Parameters

Model

Training

Session



AI

How can I help you today?



Send a message

Show controls (Ctrl+S)

Generate

Model

mistral-7b-instruct-v0.2.Q4_K_M



Load

Unload

Reload

Save settings

Model loader

llama.cpp

n-gpu-layers

Must be set to more than 0 for your GPU to be used.

0



n_ctx

Context length. Try lowering this if you run out of memory while loading the model.

32768



tensor_split

List of proportions to split the model across multiple GPUs. Example: 18,17

n_batch

512



LoRA(s)



Apply LoRAs

Whether to load the model as soon as it is selected in the Model dropdown.

Autoload the model

Download

llamacpp_HF creator

Customize instruction template

Download model or LoRA

Enter the Hugging Face username/model path, for instance: facebook/galactica-125m. To specify a branch, add it at the end after a ":" character like this: facebook/galactica-125m:main. To download a single file, enter its name in the second box.

Download

Get file list

The settings for `mistral-7b-instruct-v0.2.Q4_K_M.gguf` have been updated.

Úlohy pro velké jazykové modely

- obecné znalosti:
 - historie
 - přírodní vědy
 - technologie
 - literatura
 - aktuální události
 - zajímavosti
- vyhledávání informací:
 - vyhledávání informací o konkrétním tématu
 - shrnutí textu
 - překlad textu
 - odpovídání na otázky na základě textu
- kód a data:
 - psaní kódu v různých programovacích jazycích
 - vysvětlování dat
 - řešení problémů s kódováním
- kreativní úkoly:
 - vyprávění příběhů
 - psaní poezie
 - hudební kompozice
- další úkoly:
 - konverzace
 - organizace informací
 - řešení problémů
 - rozhodování

Nasazení modelů

- menší (s případným tréninkem)
 - klasifikace textu
- střední
 - překlad, sumarizace, ...
- větší
 - náročné úlohy

Práce s promptem

- <https://prompts.chat>
 - „Act as...“
- <https://docs.anthropic.com/claude/prompt-library>






































ANTHROPIC

User guides <> API reference **Prompt library** Search CTRL-K

Prompt library

Explore optimized prompts for a breadth of business and personal tasks.
User-submitted prompts have dark backgrounds with light colored icons (currently, there are none).
You can submit prompts via our [prompt submission form](#).

Filter: All prompts Personal prompts **Business prompts** User-submitted prompts

 Corporate clairvoyant Extract insights, identify risks, and distill key information from long corporate reports into a single memo	 Website wizard Create one-page websites based on user specifications	 Excel formula expert Create Excel formulas based on user-described calculations or data manipulations
 Google apps scripser Generate Google Apps scripts to complete tasks based on user requirements	 Python bug buster Detect and fix bugs in Python code	 Cite your sources Get answers to questions about a document's content with relevant citations supporting the response
 SQL sorcerer Transform everyday language into SQL queries	 LaTeX legend Write LaTeX documents, generating code for mathematical equations, tables, and more	 Git gud Generate appropriate Git commands based on user-described version control actions
 Meeting scribe Distill meetings into concise summaries including discussion topics, key takeaways, and action items	 Code consultant Suggest improvements to optimize Python code performance	 Function fabricator Create Python functions based on detailed specifications
 CSV converter Convert data from various formats (JSON, XML, etc.) into properly formatted CSV files	 Prose polisher Refine and improve written content with advanced copyediting techniques and suggestions	 Second grade simplifier Make complex text easy for young learners to understand
 VR fitness innovator Brainstorm creative ideas for virtual reality fitness games	 PII purifier Automatically detect and remove personally identifiable information (PII) from text documents	 Memo maestro Compose comprehensive company memos based on key points
 Career coach Engage in role-play conversations with an AI career coach	 Grading guru Compare and evaluate the quality of written texts based on user-defined criteria and standards	 Interview question crafter Generate questions for interviews
 Grammar genie Transform grammatically incorrect sentences into proper English	 Code clarifier Simplify and explain complex code in plain language	 Data organizer Turn unstructured text into bespoke JSON tables
 Brand builder Craft a design brief for a holistic brand identity	 Efficiency estimator Calculate the time complexity of functions and algorithms	 Review classifier Categorize feedback into pre-specified tags and categorizations
 Direction decoder Transform natural language into step-by-step directions	 Email extractor Extract email addresses from a document into a JSON-formatted list	 Master moderator Evaluate user inputs for potential harmful or illegal content
 Lesson planner Craft in depth lesson plans on any subject	 Product naming pro Create catchy product names from descriptions and keywords	 Spreadsheet sorcerer Generate CSV spreadsheets with various types of data
 Adaptive editor Rewrite text following user-given instructions, such as with a different tone, audience, or style	 Babel's broadcasts Create compelling product announcement tweets in the world's 10 most spoken languages	 Tweet tone detector Detect the tone and sentiment behind tweets
 Airport code analyst Find and extract airport codes from text		

Grammar genie

← Back to Prompts

Transform grammatically incorrect sentences into proper English.

Copy this prompt into our developer [Console](#) to try it for yourself!

	Content
System	Your task is to take the text provided and rewrite it into a clear, grammatically correct version while preserving the original meaning as closely as possible. Correct any spelling mistakes, punctuation errors, verb tense issues, word choice problems, and other grammatical mistakes.
User	I can haz cheeseburger?

Parametry promptu

- max_new_tokens
- temperature
- top_p
- top_k
- repetition_penalty

- seed
 - zvolené číslo – snaha o reprodukovatelnost výstupu
 - výstupy ale nejsou zaručeně deterministické

Maximální počet vygenerovaných tokenů.

Vyšší teplota způsobí, že výstupy budou náhodnější a rozmanitější.

Nižší hodnoty top-p snižují rozmanitost a zaměřují se na pravděpodobnější tokeny.

Nižší top-k také soustředí výběr vzorků na tokeny s nejvyšší pravděpodobností pro každý krok.

Trest za opakování odrazuje od opakování tokenů, které se v generovaném textu objevily nedávno.

Co má vliv na kvalitu výstupu

- počet parametrů (B)
- kontextové okno (tokeny)
- kvantizace (CPU a/nebo GPU)
 - GGUF
 - ...
- prompt: šablona
- hranice znalostí (do urč. data)
- markdown
- model loader
 - transformers
 - llama.cpp
 - ollama
 - ...

Who was Rudolf II?

gemma-2B-it (GGUF, Q8_0)

Rudolf II of Germany (1934-1945) was the last Kaiser of Germany and the longest-reigning German head of state in history. He was born in Berlin, Germany in 1934, and became Kaiser in 1937 after his father, Wilhelm II, abdicated due to health issues. During World War II, he was captured by the Allies and imprisoned in several camps, including Auschwitz. After the war, he was reinstated as Kaiser and led the country until his defeat by the Allied forces in 1945.

mistral-7B-instruct-v0.2 (GGUF, Q4_K_M)

Rudolf II (1552-1612) was the Holy Roman Emperor from 1576 until his death. He was also the King of Bohemia and Hungary during the same period. Rudolf is known for his patronage of the arts and sciences, making Prague a cultural center of Europe during his reign. He was an avid collector of art, alchemy, and rare objects. Despite his intellectual pursuits, Rudolf faced political challenges throughout his rule, including conflicts with the Protestant estates in Bohemia and Hungary, and with the Catholic Church. Ultimately, his inability to effectively govern led to his deposition by the Estates in 1611, and his successor Matthias was elected as the new Holy Roman Emperor.

Evaluace modelů

- testy
 - MT-bench (score)
 - MMLU
 - Arena Elo rating
 - ...
- leaderboard
 - <https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard>


LMSYS Chatbot Arena Leaderboard

[Vote](#) | [Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#)

LMSYS [Chatbot Arena](#) is a crowdsourced open platform for LLM evals. We've collected over **500,000** human pairwise comparisons to rank LLMs with the [Bradley-Terry model](#) and display the model ratings in Elo-scale. You can find more details in our [paper](#).

[Arena](#)
[Full Leaderboard](#)

Total #models: 82. Total #votes: 672,236. Last updated: April 13, 2024.

 **NEW!** View leaderboard for different categories (e.g., coding, long user query)!

Code to recreate leaderboard tables and plots in this [notebook](#). You can contribute your vote  at [chat.lmsys.org!](#)

Category

Overall

Overall Questions

#models: 82 (100%) #votes: 672,236 (100%)

Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo-2024-04-09	1260	+5/-5	15751	OpenAI	Proprietary	2023/12
1	Claude 3 Opus	1255	+3/-4	56101	Anthropic	Proprietary	2023/8
1	GPT-4-1106-preview	1254	+3/-3	65159	OpenAI	Proprietary	2023/4
2	GPT-4-0125-preview	1250	+3/-4	50923	OpenAI	Proprietary	2023/12
5	Bard (Gemini Pro)	1209	+5/-5	12468	Google	Proprietary	Online
5	Claude 3 Sonnet	1203	+3/-3	62056	Anthropic	Proprietary	2023/8
7	Command R+	1193	+4/-4	29437	Cohere	CC-BY-NC-4.0	2024/3
7	GPT-4-0314	1189	+4/-4	42925	OpenAI	Proprietary	2021/9
9	Claude 3 Haiku	1182	+3/-3	57727	Anthropic	Proprietary	2023/8
10	GPT-4-0613	1164	+3/-3	61520	OpenAI	Proprietary	2021/9
10	Mistral Large 2402	1158	+3/-4	37650	Mistral	Proprietary	Unknown

Rozšířené využití modelů

- asistenti
 - GPT-4 assistant
 - text-generation-webui character
- práce s vlastními daty (retrieval augmented generation, RAG)
 - GPT-4
 - localGPT
- práce se soubory
 - jako předchozí + překlady atd.
- přístup přes API
 - OpenAI API (lze i emulovat)