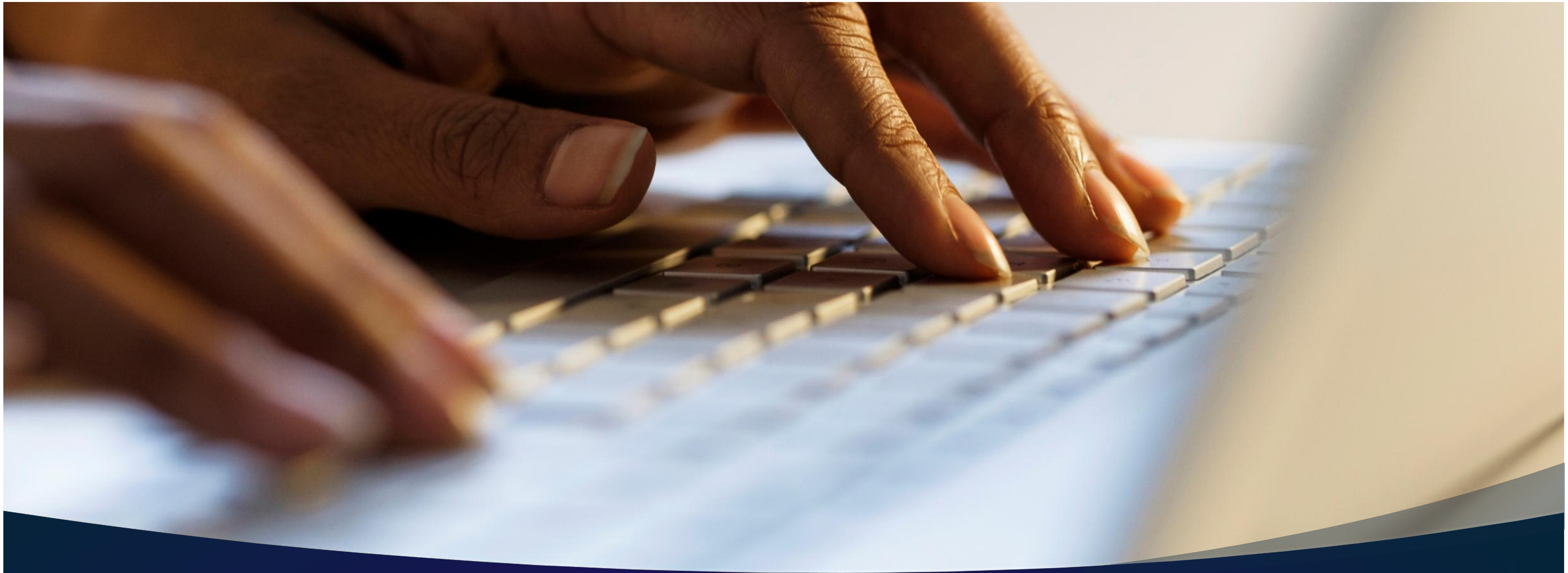


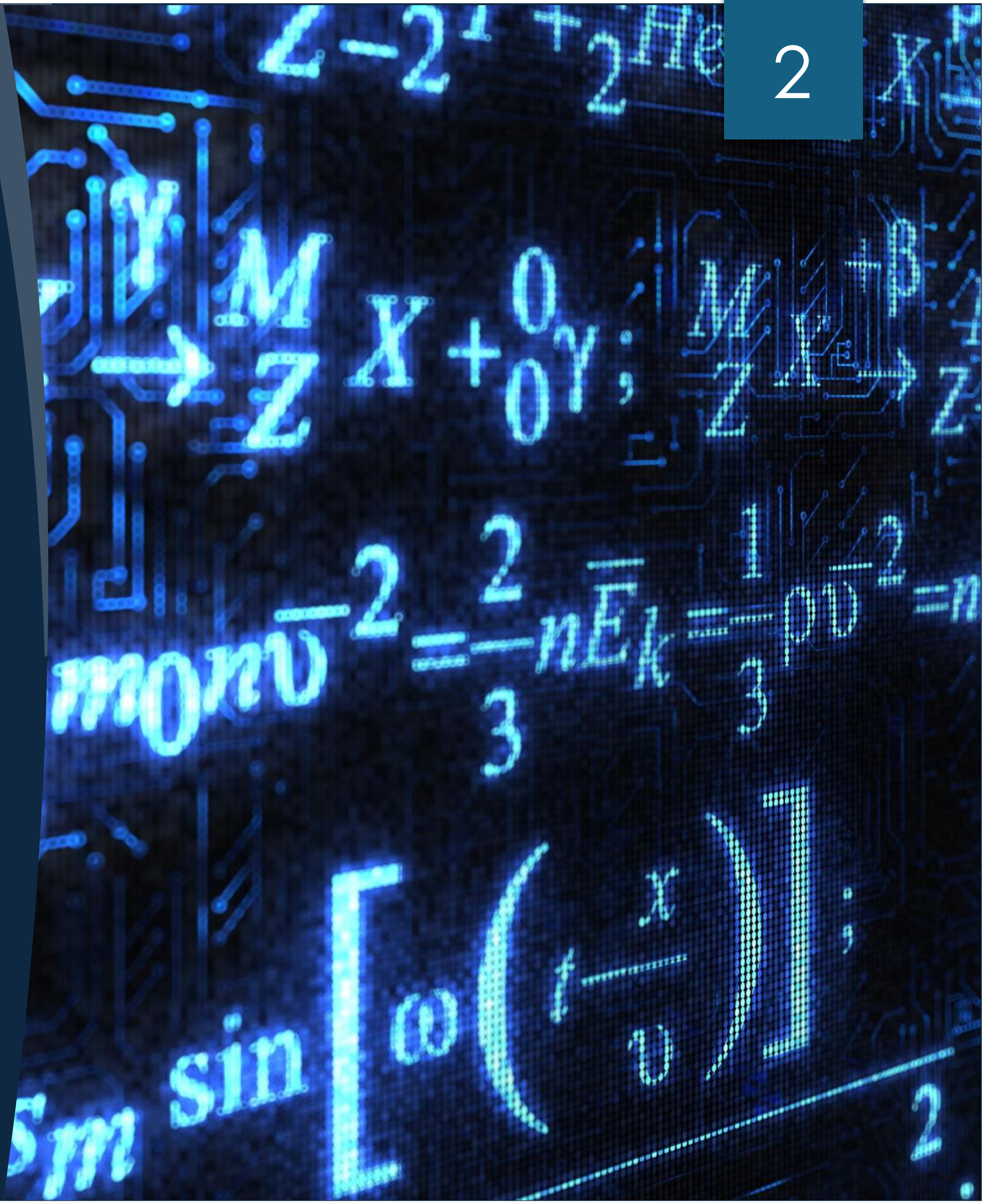
Keeping Things Local

Making Your Own Private LLM

CORVUS | BRONWEN AKER

What We Will Explore

- Brief Overview of AI & LLMs
- Why Keep an LLM Local?
  - Security & Ethical Concerns
- Popular Local LLM Options
- Ways to Enhance Your LLM
- Demo:
  - Ollama Set-Up, Use, Customization



Why Should You Care What I Say?

- Bronwen Aker | Corvus | The Cybrarian
- 30+ years development experience
  - Web, desktop app, mobile app, etc.
- Experienced Technical Trainer
- Switched to Cybersecurity in 2017
- Technical Editor for BHIS since 2018
- Latest Obsession: AI Research
- Bottomline: I’m a geek who has been around and seen a lot of 💩

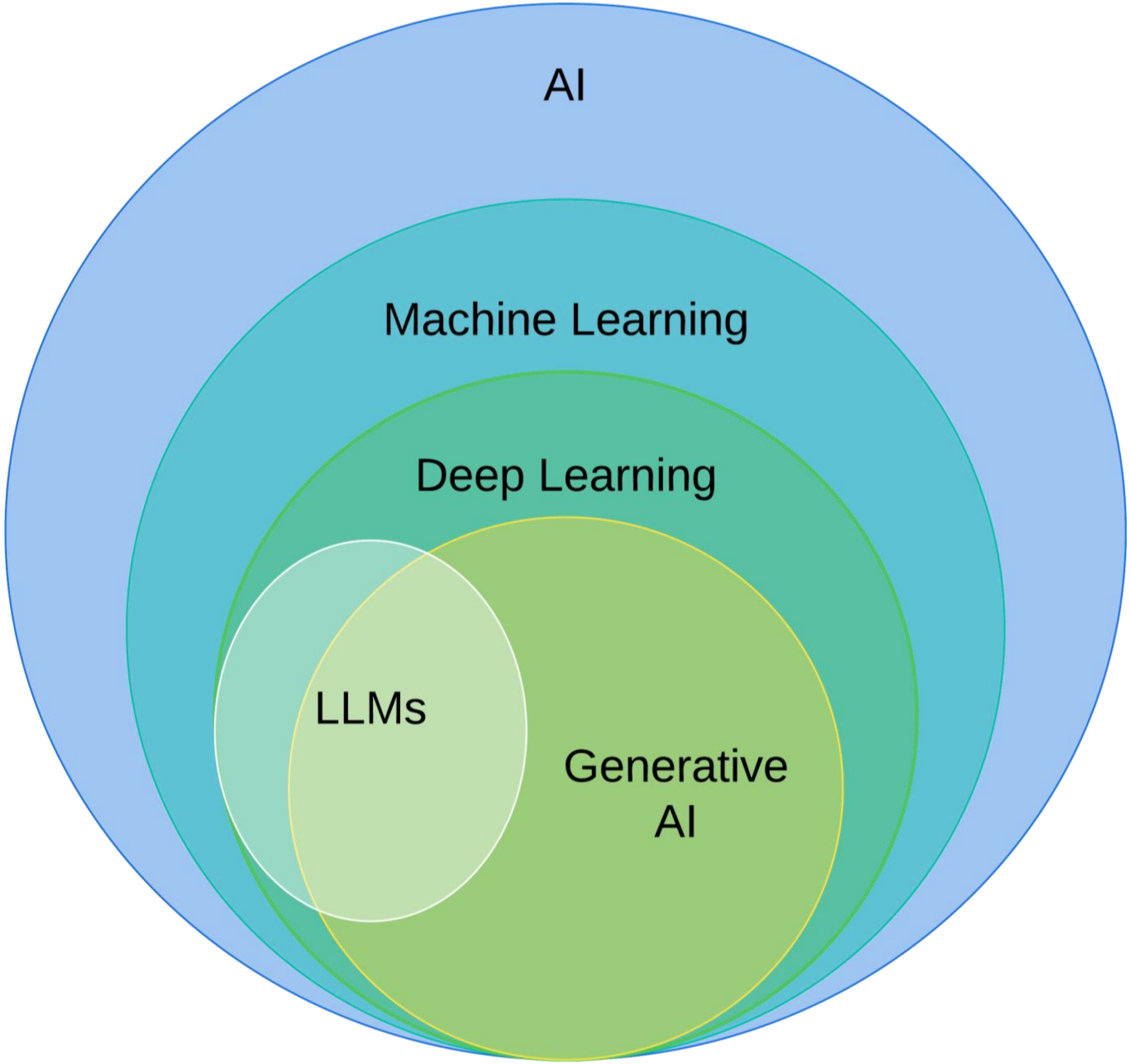
FULL DISCLOSURE

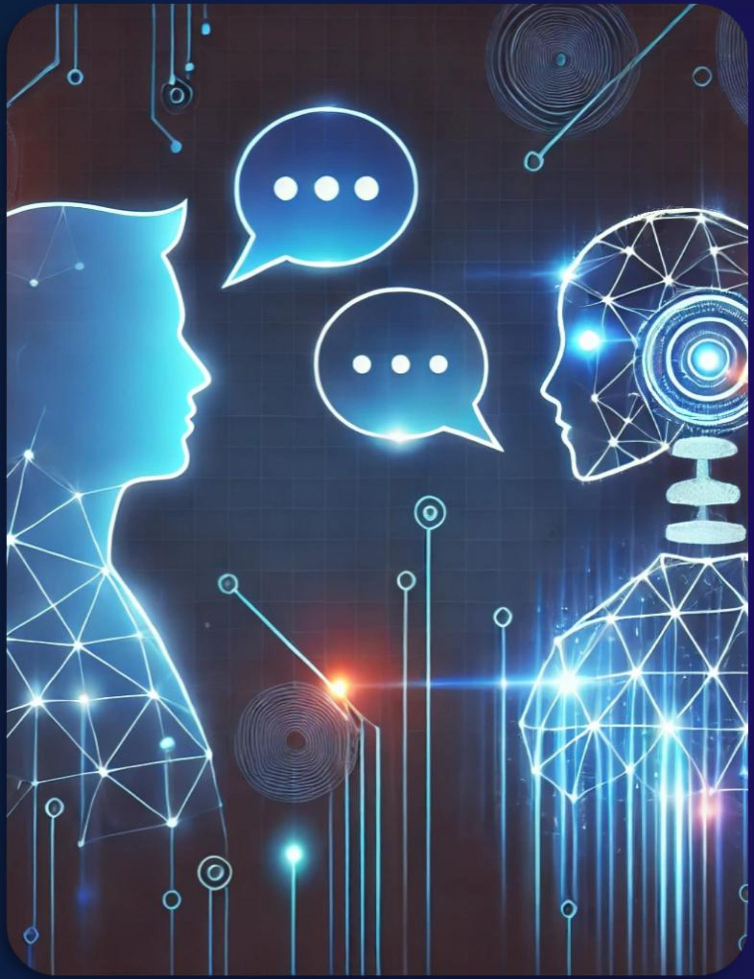
- This presentation was created using AIs like ChatGPT, Copilot, DALL-E, and Midjourney
- They were, at all times, under adult supervision*

* Assuming that you consider me to be an adult. 😉

What is AI?

- AI is a vast computer science branch aimed at creating systems that perform tasks requiring human intelligence
- AI includes:
  - Robotics
  - Computer Vision
  - Natural Language Processing (NLP)
  - Expert Systems





Where Do LLMs Fit in AI?

- LLMs are a subset of generative AI
- Text based
- Probabilistic
- Current models trained on MASSIVE amounts of data
- Most are general purpose
- Easy access to LLMs has popularized AI

Why Go Local?

- Privacy & Security
  - No data leaves your machine/network/intranet
  - Sensitive files/data remains in YOUR control
- Customization
  - Fine-tune for your needs
    - Create custom models
    - Use Retrieval-Augmented Generation (RAG)
    - Train your own model (DEEP Rabbit Hole!)

Cybersecurity Concerns About LLMs

- Jailbreaking & Prompt Injection
- Data Leakage & Privacy Risks
- Model Bias & Poisoning
- Social Engineering Automation
- Poor API Implementation & Authentication Controls

Components to Build a Local LLM

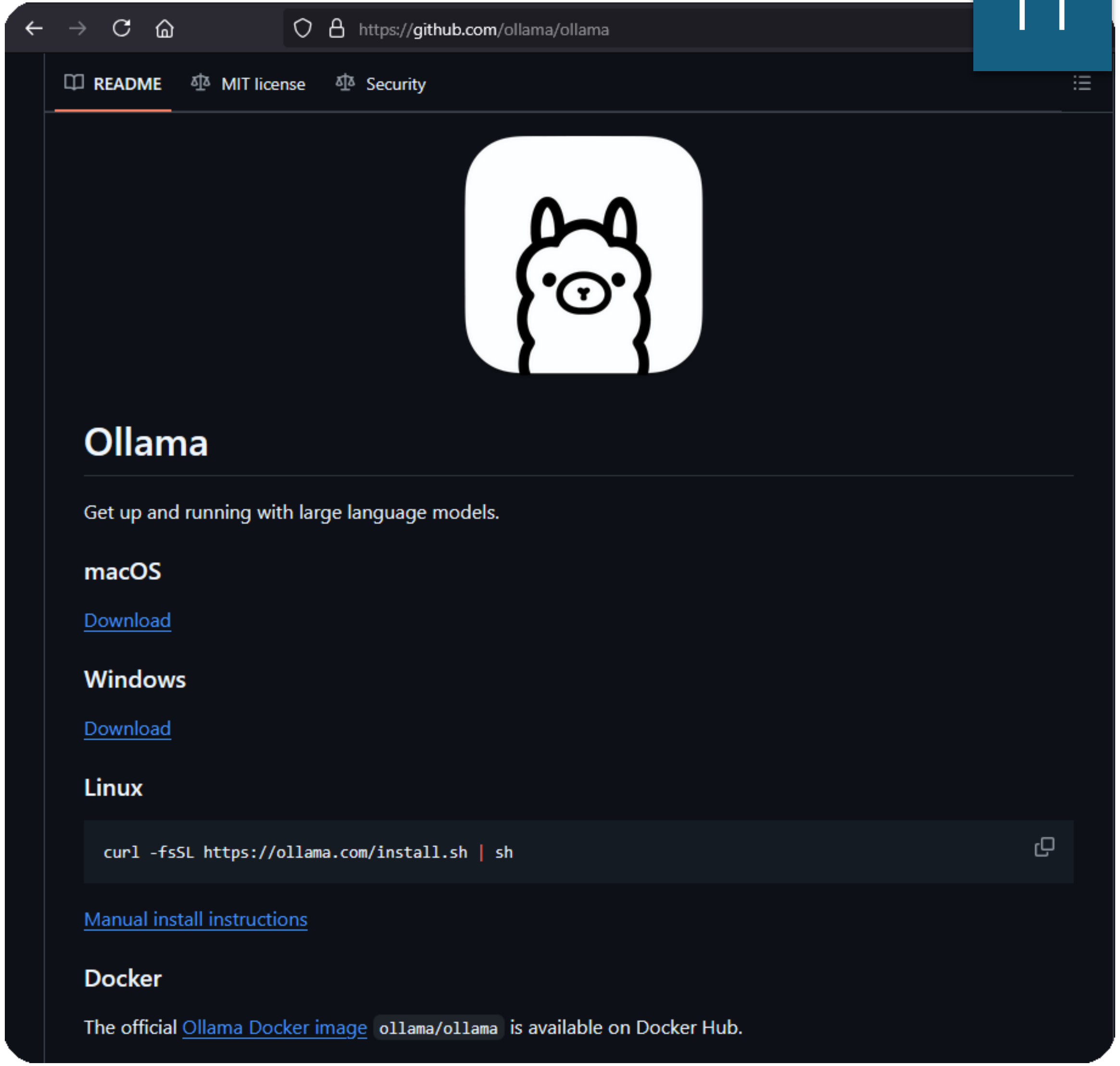
- Hardware
  - GPUs speed processing, but not 100% necessary
  - RAM & Disk Space: 16GB+ RAM, 50GB+ disk space
- Software
  - Determines UI, model management, other capabilities
  - Ollama, LM Studio, GPT4All
- LLM Model
  - Open-source models available from Hugging Face, Mistral, LLaMA, and others
  - Llama 3.3, DeepSeek-R1, Phi-4, Mistral, Gemma 2

GPUs

- NVIDIA – Best for AI workloads
  - RTX 4090 (24GB VRAM) – High-end consumer option
  - RTX 3090/3090 Ti (24GB VRAM) – Older but powerful
  - A100 (40GB/80GB) – Enterprise-grade, excellent for large models
  - H100 (80GB) – Top-tier but very expensive
- AMD – Limited AI support (less optimized than NVIDIA)
  - Radeon RX 7900 XTX (24GB VRAM) – High VRAM, but uses ROCm, not CUDA

Software: Ollama

- Command Line Interface (CLI)
- Open Source
  - MIT License
- Useable on:
  - Linux, Mac, Winderz, Docker
- Active communities on Discord & Reddit
- Pulls models from Ollama repo



Ollama

Get up and running with large language models.

macOS

Download

Windows

Download

Linux

```
curl -fsSL https://ollama.com/install.sh | sh
```

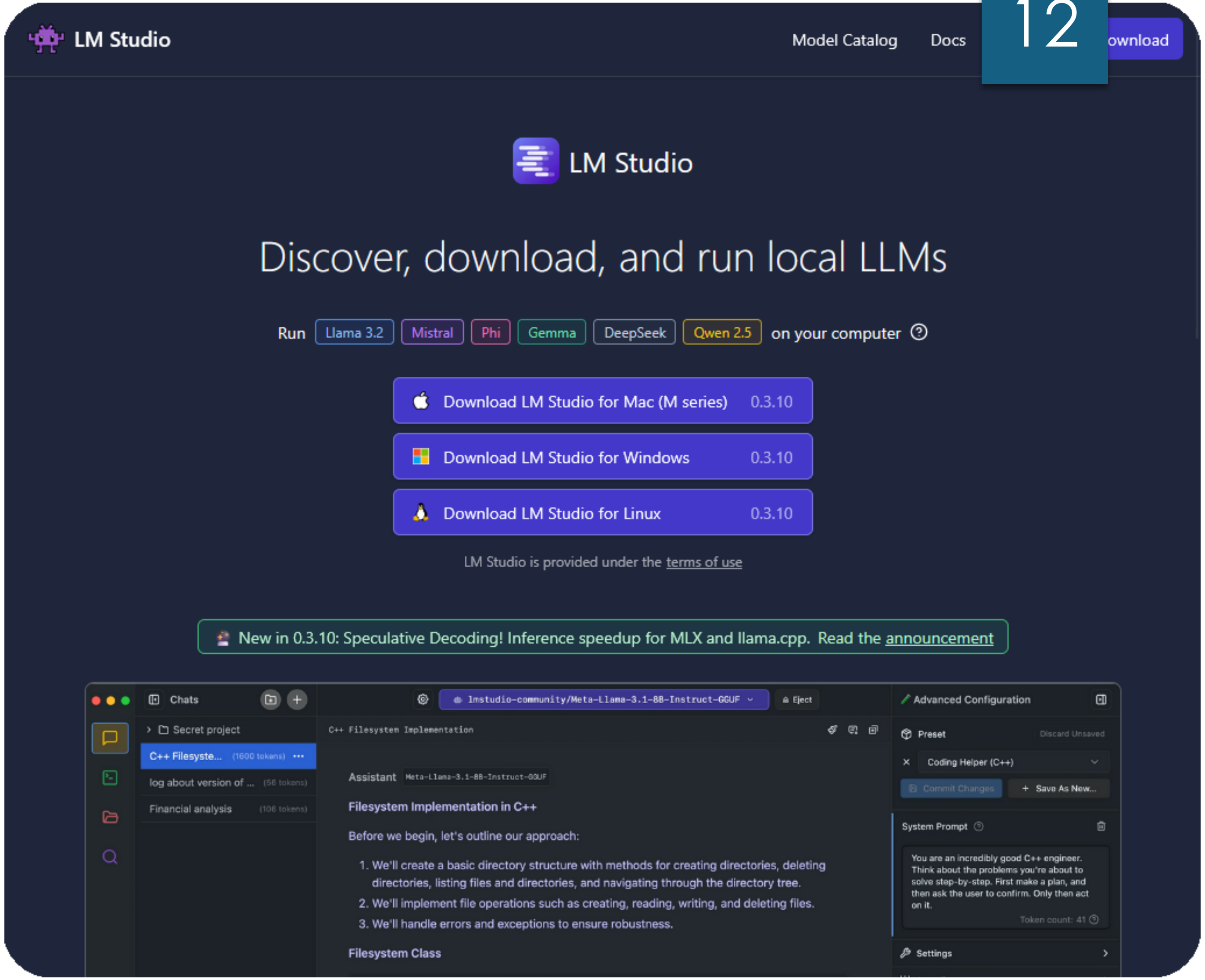
Manual install instructions

Docker

The official Ollama Docker image `ollama/ollama` is available on Docker Hub.

Software: LM Studio

- GUI and CLI
- Developed by Element Labs, Inc.
  - Limited license for personal, non-commercial use
- Useable on:
  - Linux, Mac, Winderz
- Active community on Discord
- Pulls models from Hugging Face
- Native RAG features


LM Studio
Model Catalog
Docs
ownload
LM Studio

Discover, download, and run local LLMs

Run Llama 3.2 Mistral Phi Gemma DeepSeek Qwen 2.5 on your computer

Download LM Studio for Mac (M series) 0.3.10
Download LM Studio for Windows 0.3.10
Download LM Studio for Linux 0.3.10

LM Studio is provided under the terms of use

New in 0.3.10: Speculative Decoding! Inference speedup for MLX and llama.cpp. Read the announcement

Chats
Secret project
C++ Filesyste... (1600 tokens)
log about version of ... (56 tokens)
Financial analysis (106 tokens)
lmstudio-community/Meta-Llama-3.1-8B-Instruct-GGUF
Eject
C++ Filesystem Implementation
Advanced Configuration
Preset
Discard Unsaved
Coding Helper (C++)
Commit Changes
+ Save As New...
System Prompt
Settings

Assistant Meta-Llama-3.1-8B-Instruct-GGUF

Filesystem Implementation in C++

Before we begin, let's outline our approach:

1. We'll create a basic directory structure with methods for creating directories, deleting directories, listing files and directories, and navigating through the directory tree.
2. We'll implement file operations such as creating, reading, writing, and deleting files.
3. We'll handle errors and exceptions to ensure robustness.

Filesystem Class

You are an incredibly good C++ engineer. Think about the problems you're about to solve step-by-step. First make a plan, and then ask the user to confirm. Only then act on it.

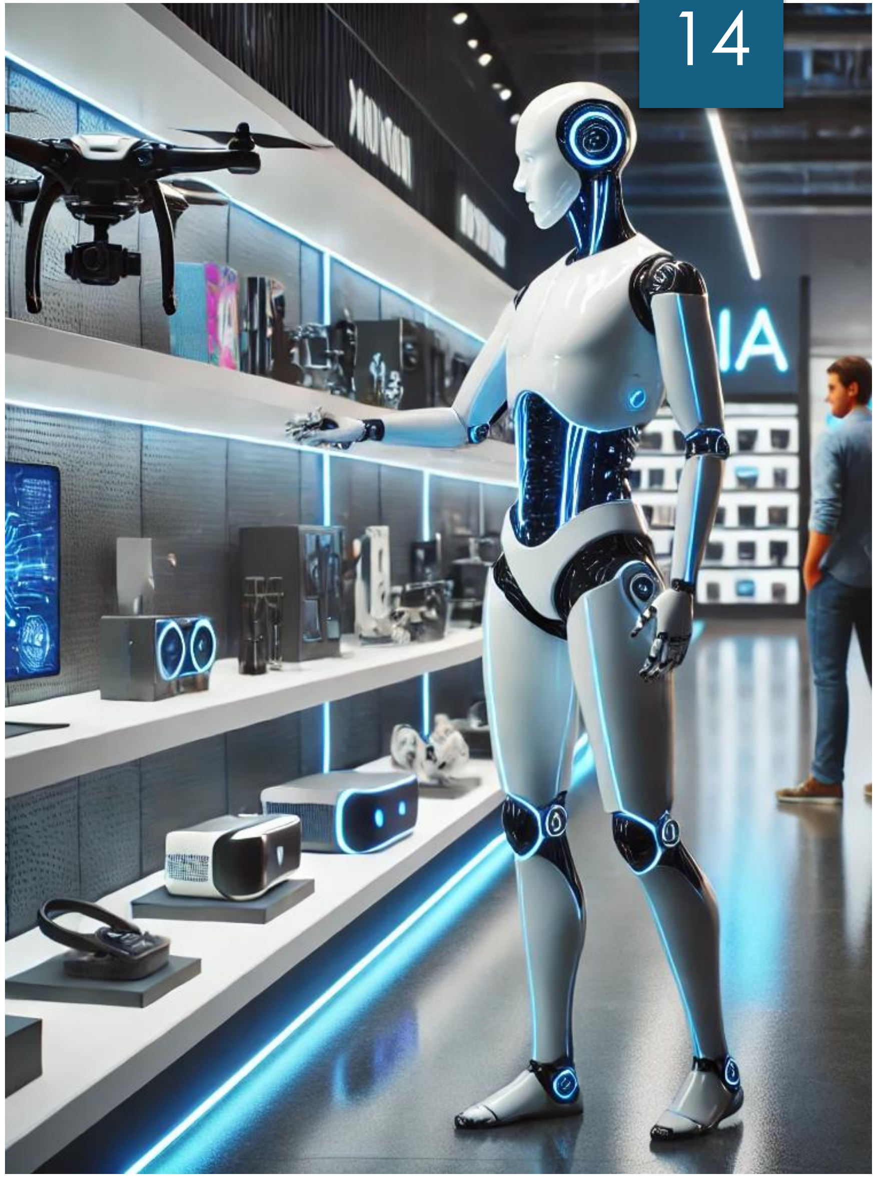
Token count: 41

Many Other Software Options

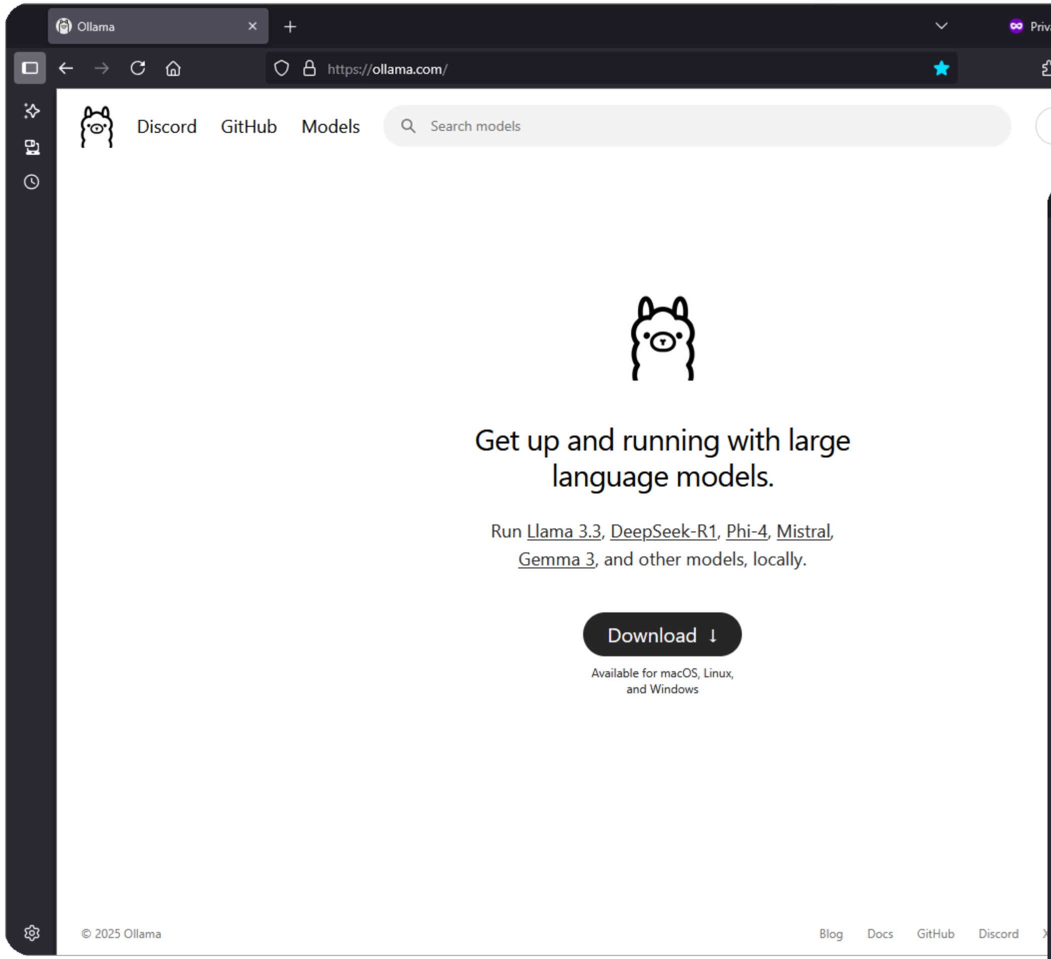
- Backyard AI
- gpt4all
- Jan
- Jellybox
- llama.cpp
- LocalAI
- Msty
- node-llama-cpp
- RecurseChat
- Sanctum
- TGI
- vLLM

Choosing an LLM Model

- LLM Model Naming Conventions
- Base Name = Core model name
- Version Number
- Parameter Count: billions (B) or trillions (T) of parameters
- Training Type
  - Base, Chat, Instruct
- Additional Modifiers
- For example, **Mistral 7B Instruct** means:
  - Developed by Mistral AI
  - 7 billion parameters
  - Fine-tuned for following instructions



Ollama & Hugging Face

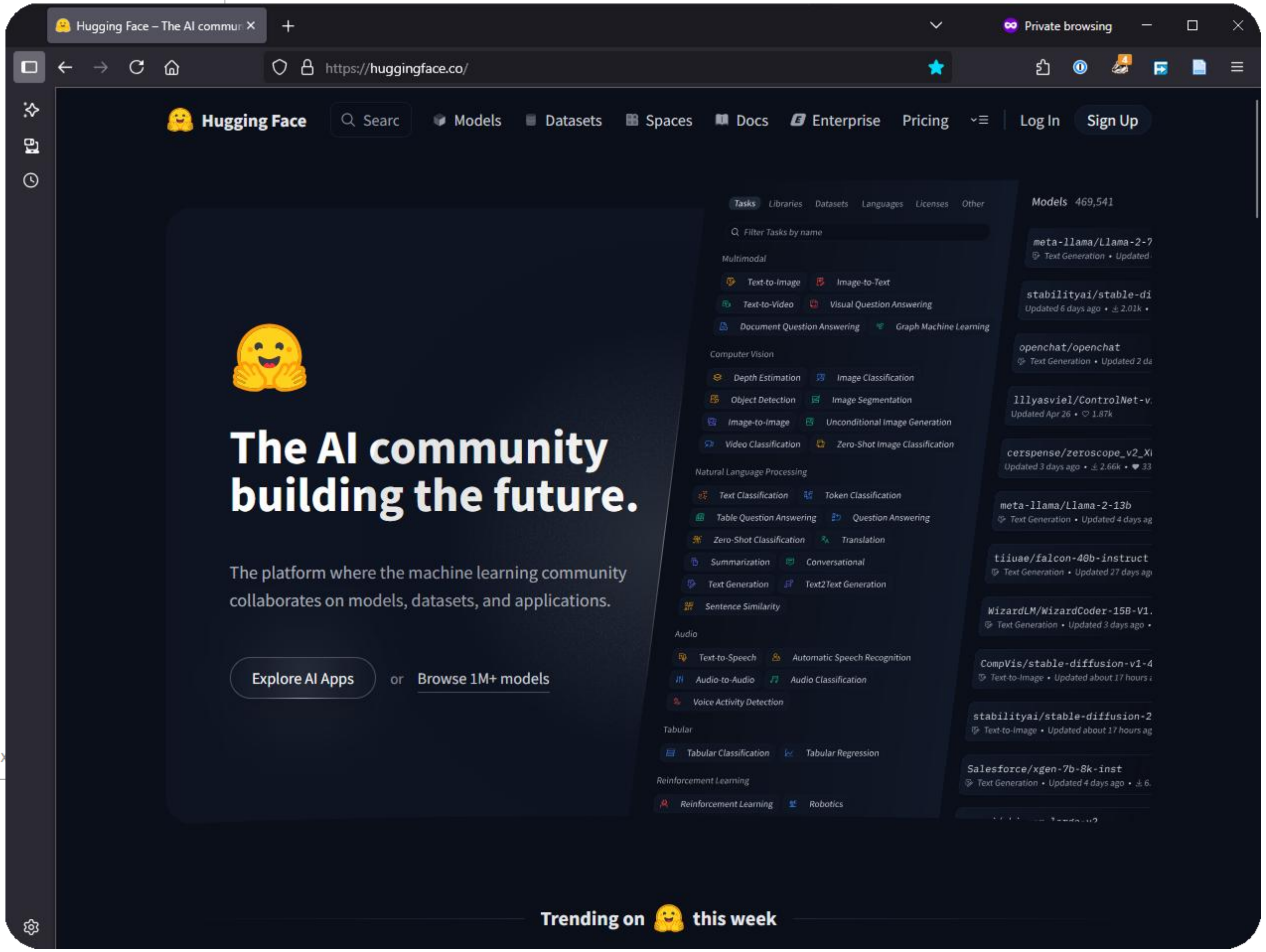


Get up and running with large language models.

Run Llama 3.3, DeepSeek-R1, Phi-4, Mistral, Gemma 3, and other models, locally.

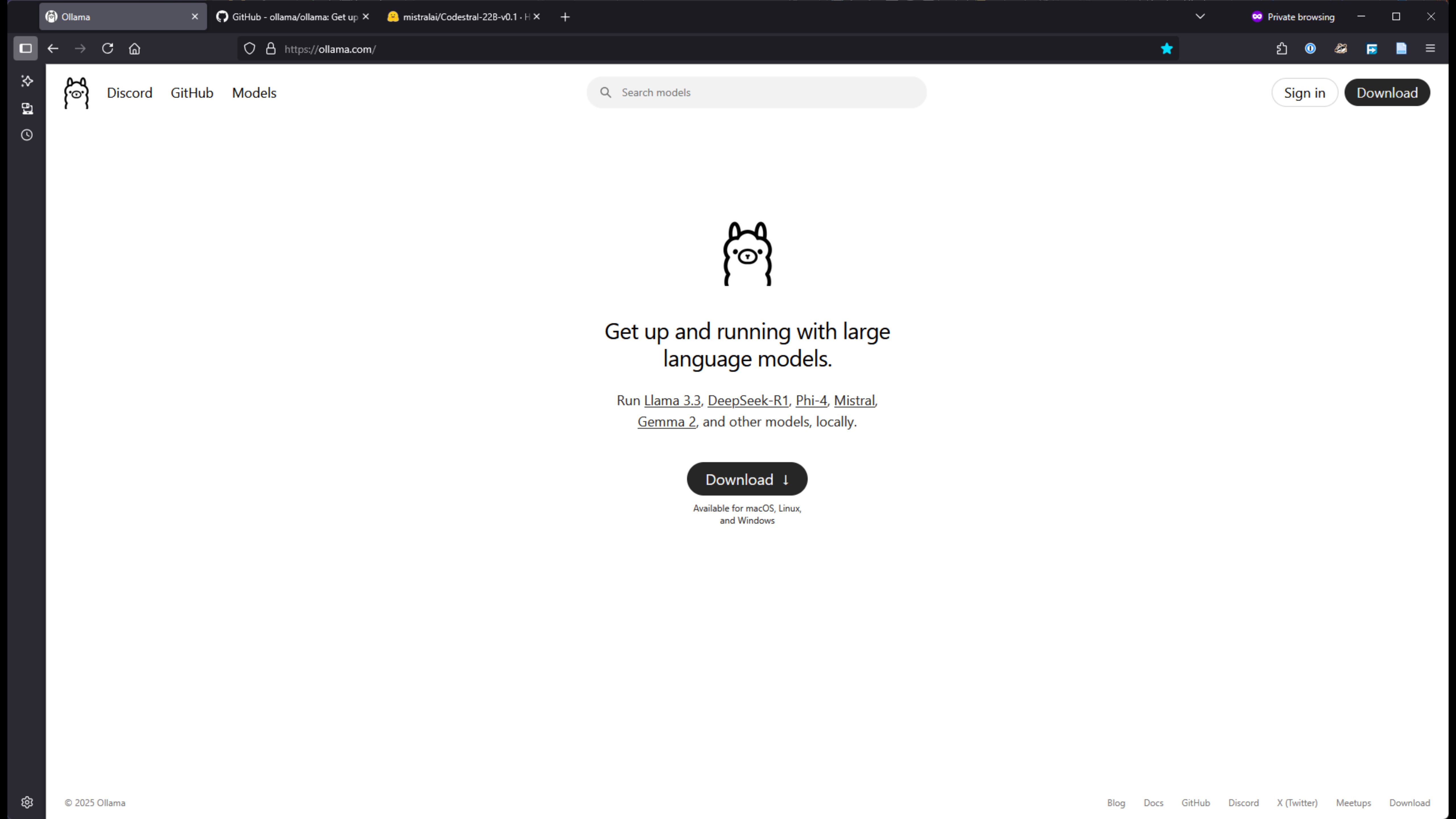
Available for macOS, Linux, and Windows

© 2025 Ollama



The AI community building the future.

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

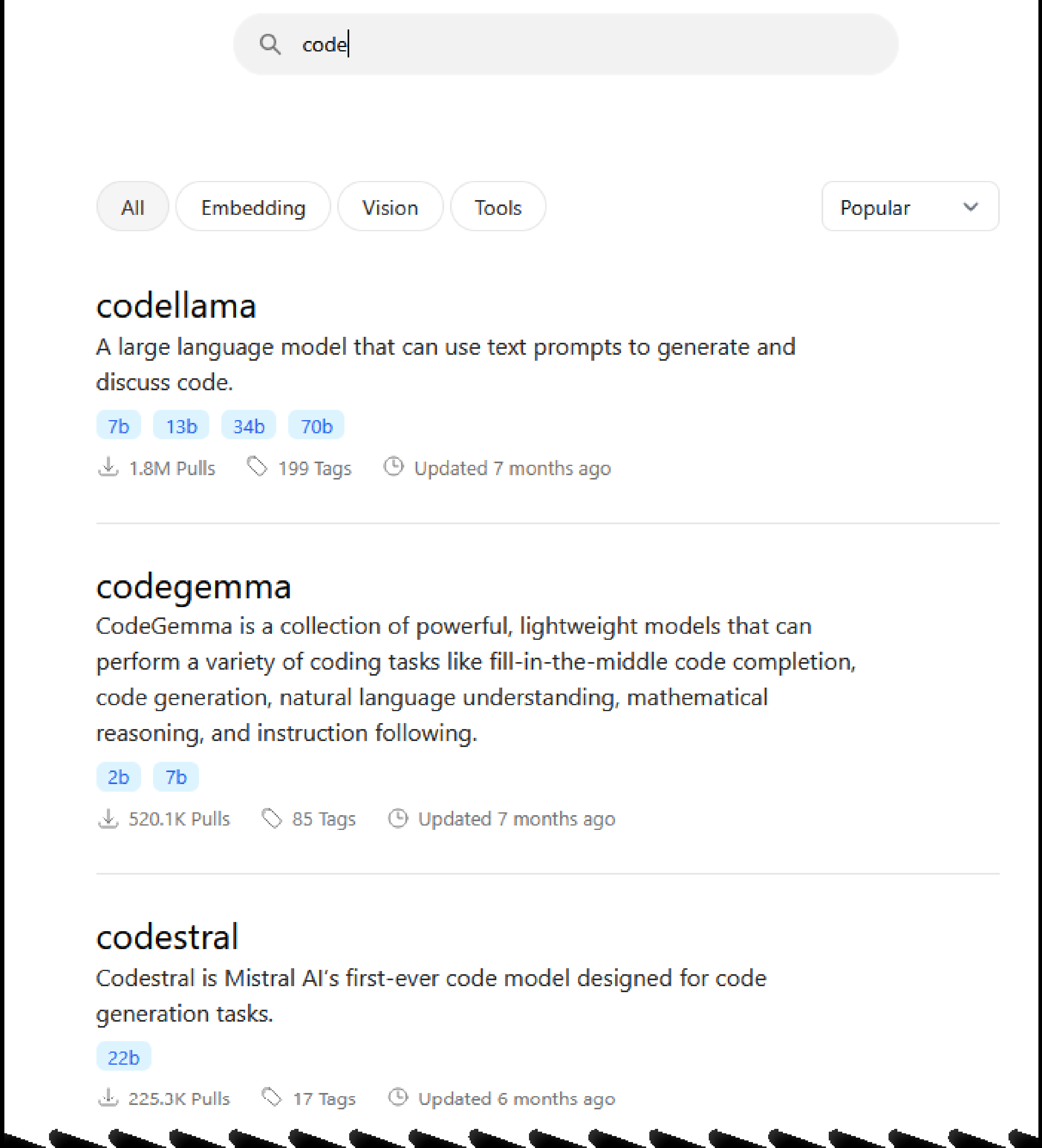


Get up and running with large language models.

Run Llama 3.3, DeepSeek-R1, Phi-4, Mistral, Gemma 2, and other models, locally.

Available for macOS, Linux, and Windows

© 2025 Ollama



code

codellama

A large language model that can use text prompts to generate and discuss code.

7b 13b 34b 70b

1.8M Pulls 199 Tags Updated 7 months ago

codegemma

CodeGemma is a collection of powerful, lightweight models that can perform a variety of coding tasks like fill-in-the-middle code completion, code generation, natural language understanding, mathematical reasoning, and instruction following.

2b 7b

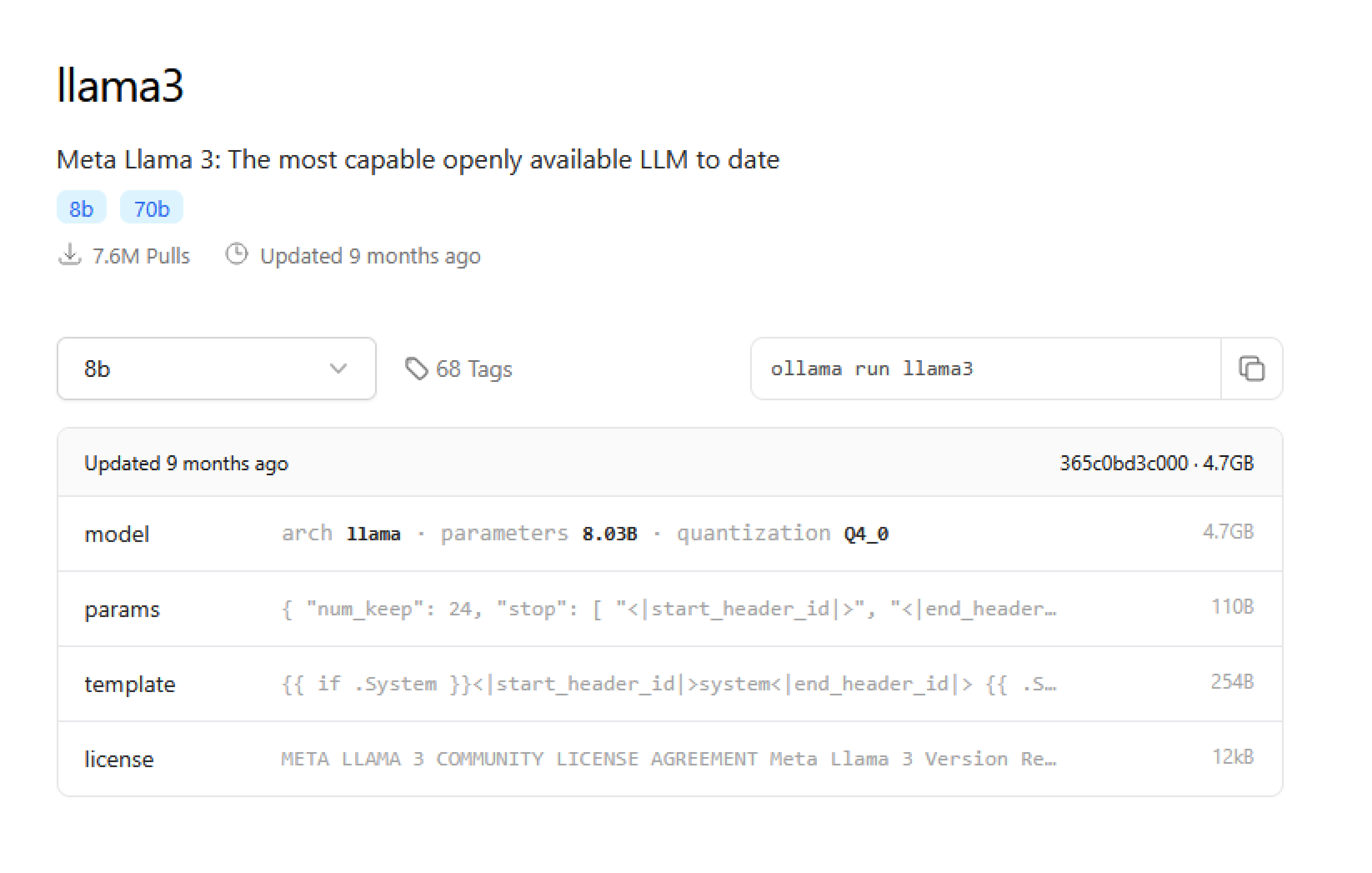
520.1K Pulls 85 Tags Updated 7 months ago

codestral

Codestral is Mistral AI's first-ever code model designed for code generation tasks.

22b

225.3K Pulls 17 Tags Updated 6 months ago



llama3

Meta Llama 3: The most capable openly available LLM to date

8b 70b

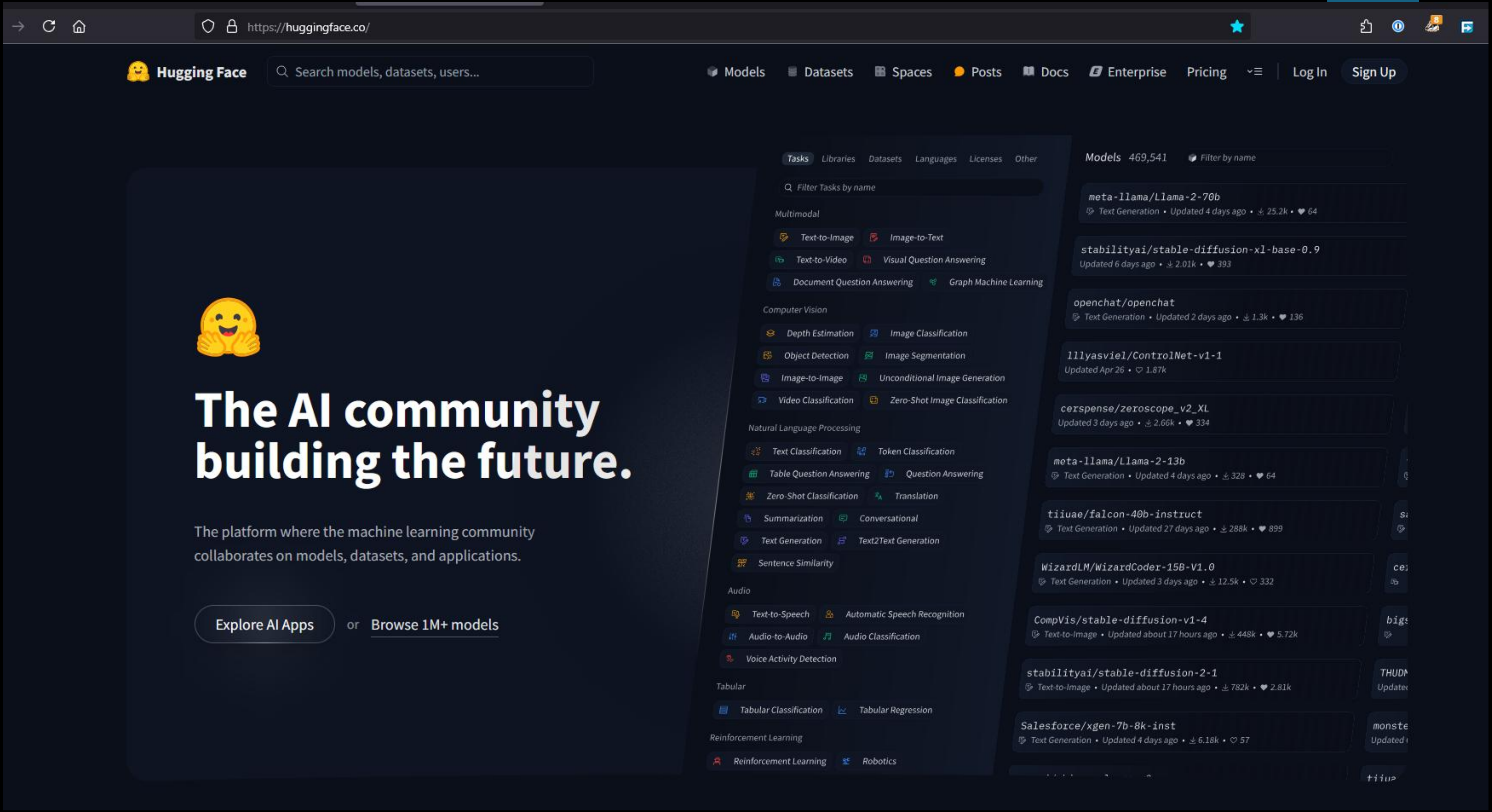
7.6M Pulls Updated 9 months ago

8b 68 Tags

```
ollama run llama3
```

Updated 9 months ago 365c0bd3c000 · 4.7GB

model	arch **llama** · parameters **8.03B** · quantization **Q4_0**	4.7GB
params	`{ "num_keep": 24, "stop": [ "<|start_header_id|>", "<|end_header…`	110B
template	`{{ if .System }}<|start_header_id|>system<|end_header_id|> {{ .S…`	254B
license	META LLAMA 3 COMMUNITY LICENSE AGREEMENT Meta Llama 3 Version Re…	12kB



The AI community building the future.

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

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

lllyasviel/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

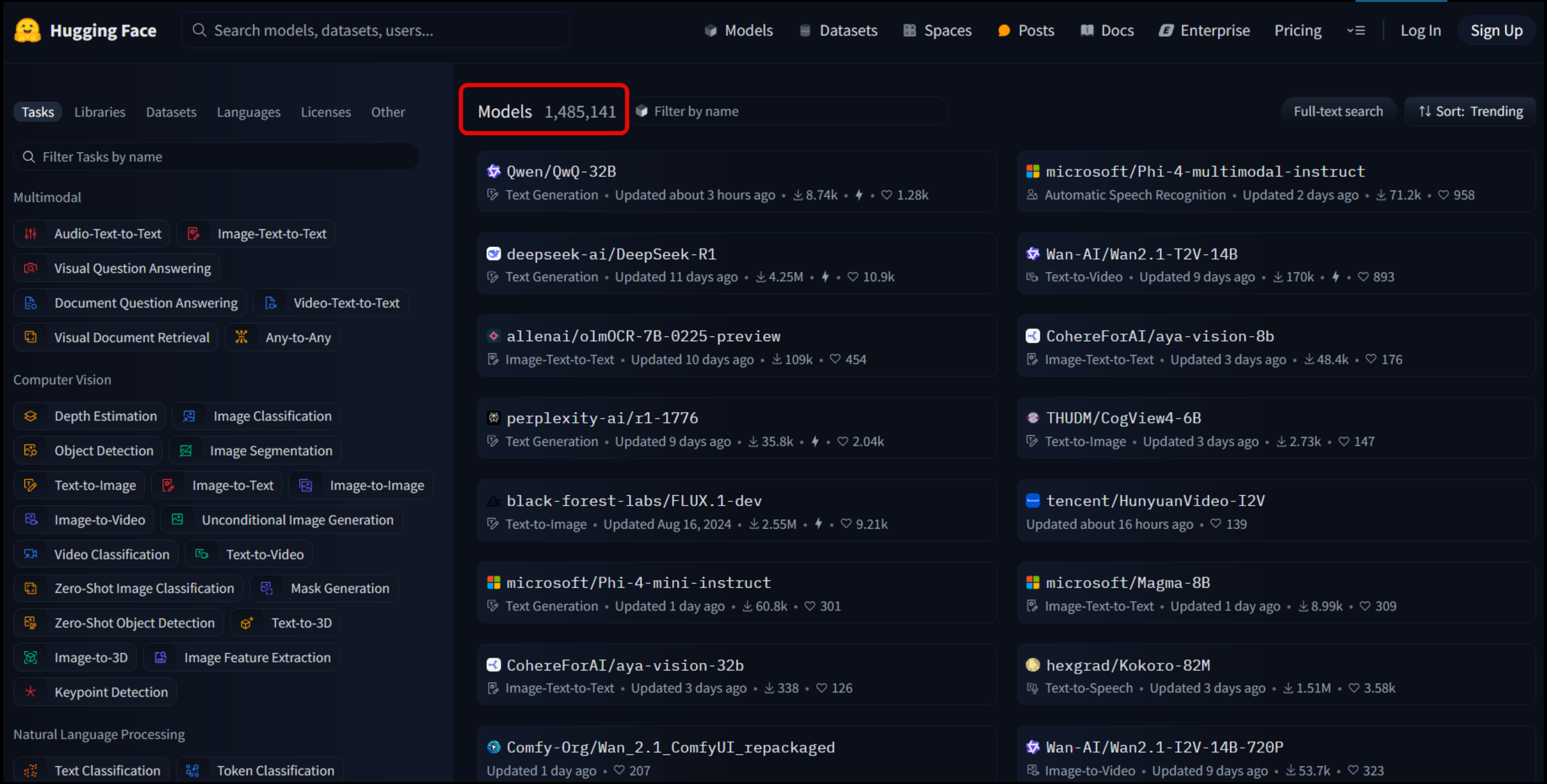
tiiuae/falcon-40b-instruct
Text Generation • Updated 27 days ago • 288k • 899

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



Models 1,485,141

Multimodal

- Audio-Text-to-Text
- Image-Text-to-Text
- Visual Question Answering
- Document Question Answering
- Video-Text-to-Text
- Visual Document Retrieval
- Any-to-Any

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
- Zero-Shot Image Classification
- Mask Generation
- Zero-Shot Object Detection
- Text-to-3D
- Image-to-3D
- Image Feature Extraction
- Keypoint Detection

Natural Language Processing

- Text Classification
- Token Classification

Qwen/QwQ-32B
Text Generation • Updated about 3 hours ago • 8.74k • ϟ • ♡ 1.28k

microsoft/Phi-4-multimodal-instruct
Automatic Speech Recognition • Updated 2 days ago • 71.2k • ♡ 958

deepseek-ai/DeepSeek-R1
Text Generation • Updated 11 days ago • 4.25M • ϟ • ♡ 10.9k

Wan-AI/Wan2.1-T2V-14B
Text-to-Video • Updated 9 days ago • 170k • ϟ • ♡ 893

allenai/olmOCR-7B-0225-preview
Image-Text-to-Text • Updated 10 days ago • 109k • ♡ 454

CohereForAI/aya-vision-8b
Image-Text-to-Text • Updated 3 days ago • 48.4k • ♡ 176

perplexity-ai/r1-1776
Text Generation • Updated 9 days ago • 35.8k • ϟ • ♡ 2.04k

THUDM/CogView4-6B
Text-to-Image • Updated 3 days ago • 2.73k • ♡ 147

black-forest-labs/FLUX.1-dev
Text-to-Image • Updated Aug 16, 2024 • 2.55M • ϟ • ♡ 9.21k

tencent/HunyuanVideo-I2V
Updated about 16 hours ago • ♡ 139

microsoft/Phi-4-mini-instruct
Text Generation • Updated 1 day ago • 60.8k • ♡ 301

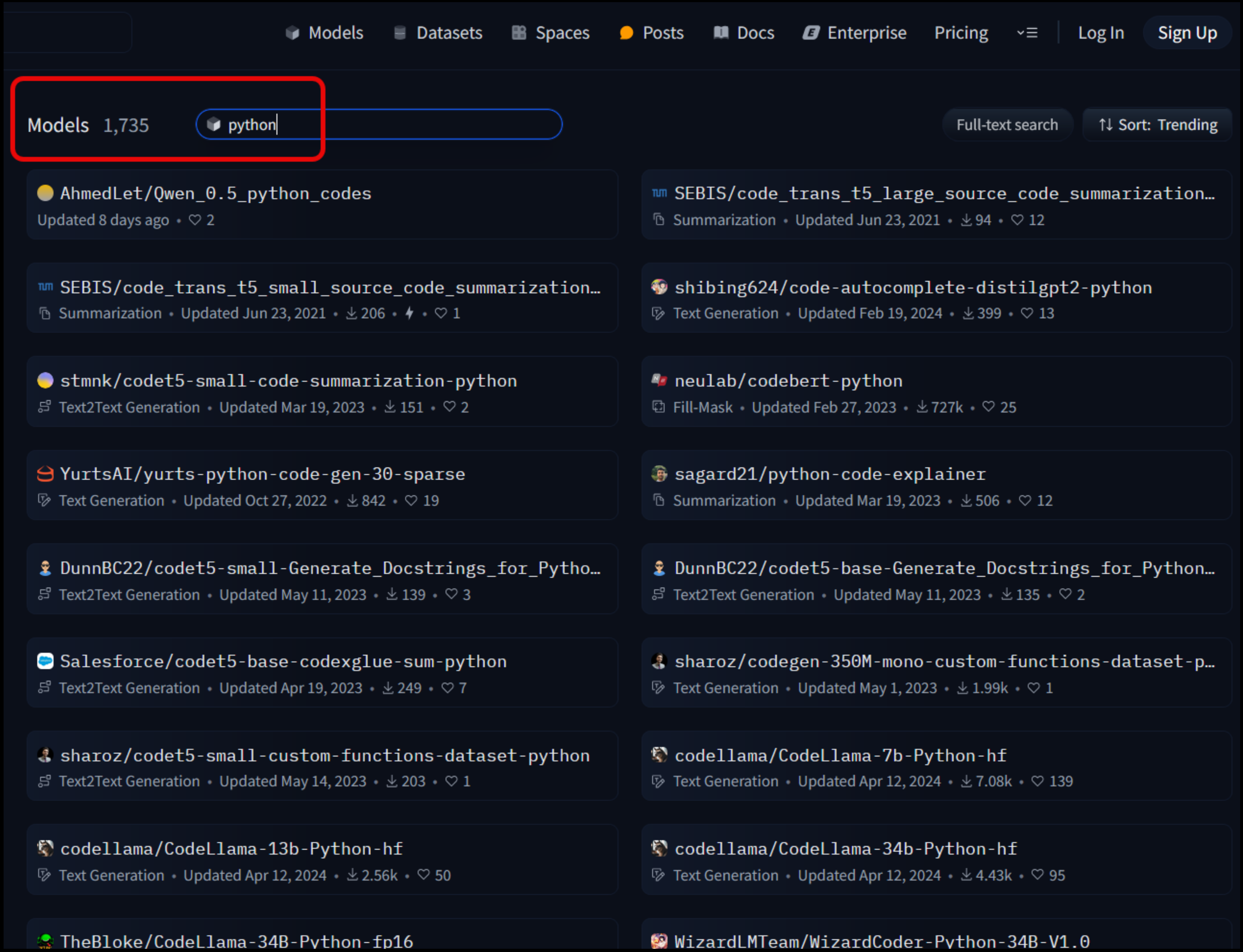
microsoft/Magma-8B
Image-Text-to-Text • Updated 1 day ago • 8.99k • ♡ 309

CohereForAI/aya-vision-32b
Image-Text-to-Text • Updated 3 days ago • 338 • ♡ 126

hexgrad/Kokoro-82M
Text-to-Speech • Updated 3 days ago • 1.51M • ♡ 3.58k

Comfy-Org/Wan_2.1_ComfyUI_repackaged
Updated 1 day ago • ♡ 207

Wan-AI/Wan2.1-I2V-14B-720P
Image-to-Video • Updated 9 days ago • 53.7k • ♡ 323



Models 1,735 python

AhmedLet/Qwen_0.5_python_codes
Updated 8 days ago • ♡ 2

SEBIS/code_trans_t5_large_source_code_summarization…
Summarization • Updated Jun 23, 2021 • ⤓ 94 • ♡ 12

SEBIS/code_trans_t5_small_source_code_summarization…
Summarization • Updated Jun 23, 2021 • ⤓ 206 • ϟ • ♡ 1

shibing624/code-autocomplete-distilgpt2-python
Text Generation • Updated Feb 19, 2024 • ⤓ 399 • ♡ 13

stmnk/codet5-small-code-summarization-python
Text2Text Generation • Updated Mar 19, 2023 • ⤓ 151 • ♡ 2

neulab/codebert-python
Fill-Mask • Updated Feb 27, 2023 • ⤓ 727k • ♡ 25

YurtsAI/yurts-python-code-gen-30-sparse
Text Generation • Updated Oct 27, 2022 • ⤓ 842 • ♡ 19

sagard21/python-code-explainer
Summarization • Updated Mar 19, 2023 • ⤓ 506 • ♡ 12

DunnBC22/codet5-small-Generate_Docstrings_for_Pytho…
Text2Text Generation • Updated May 11, 2023 • ⤓ 139 • ♡ 3

DunnBC22/codet5-base-Generate_Docstrings_for_Python…
Text2Text Generation • Updated May 11, 2023 • ⤓ 135 • ♡ 2

Salesforce/codet5-base-codexglue-sum-python
Text2Text Generation • Updated Apr 19, 2023 • ⤓ 249 • ♡ 7

sharoz/codegen-350M-mono-custom-functions-dataset-p…
Text Generation • Updated May 1, 2023 • ⤓ 1.99k • ♡ 1

sharoz/codet5-small-custom-functions-dataset-python
Text2Text Generation • Updated May 14, 2023 • ⤓ 203 • ♡ 1

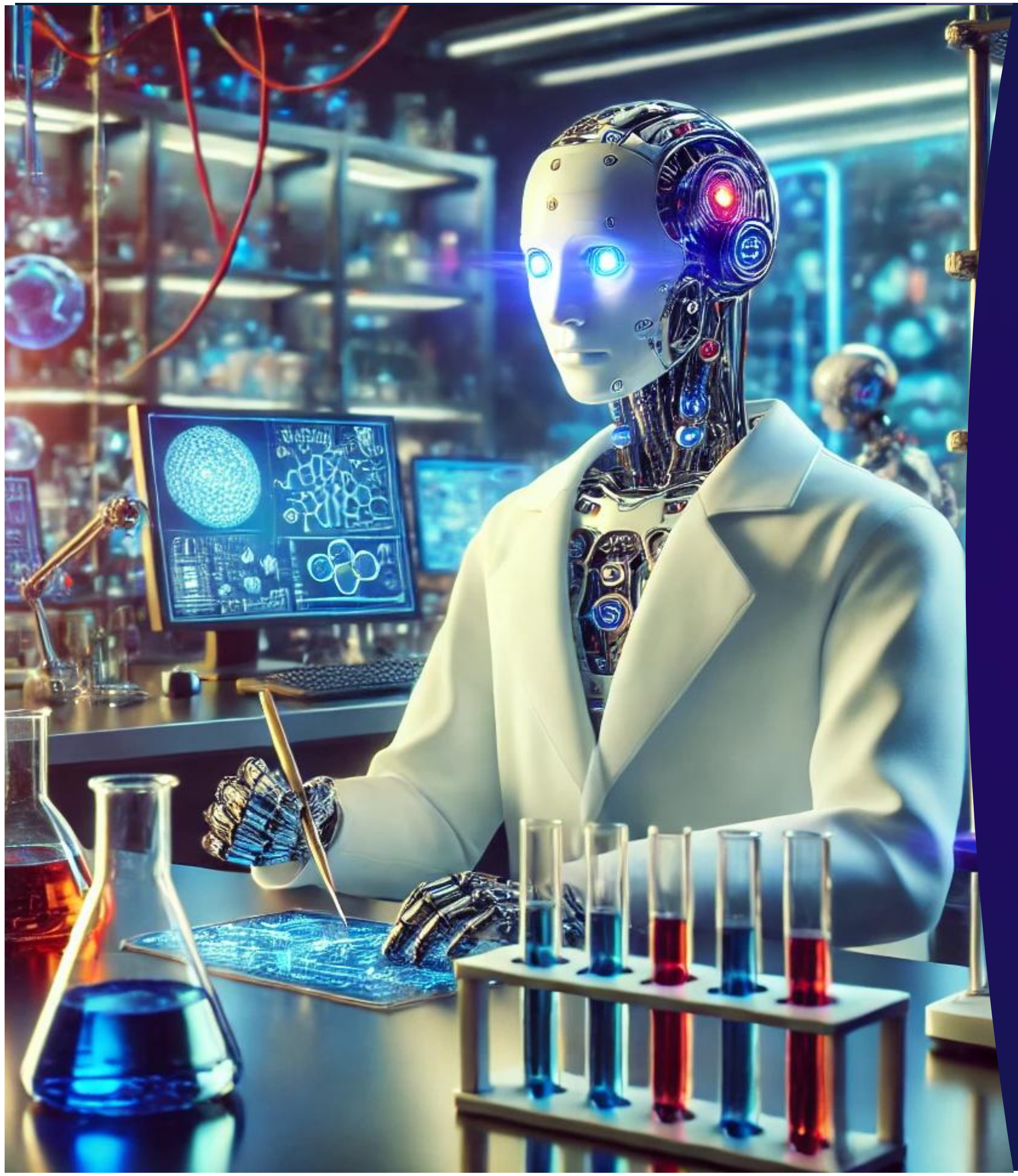
codellama/CodeLlama-7b-Python-hf
Text Generation • Updated Apr 12, 2024 • ⤓ 7.08k • ♡ 139

codellama/CodeLlama-13b-Python-hf
Text Generation • Updated Apr 12, 2024 • ⤓ 2.56k • ♡ 50

codellama/CodeLlama-34b-Python-hf
Text Generation • Updated Apr 12, 2024 • ⤓ 4.43k • ♡ 95

TheBloke/CodeLlama-34B-Python-fp16

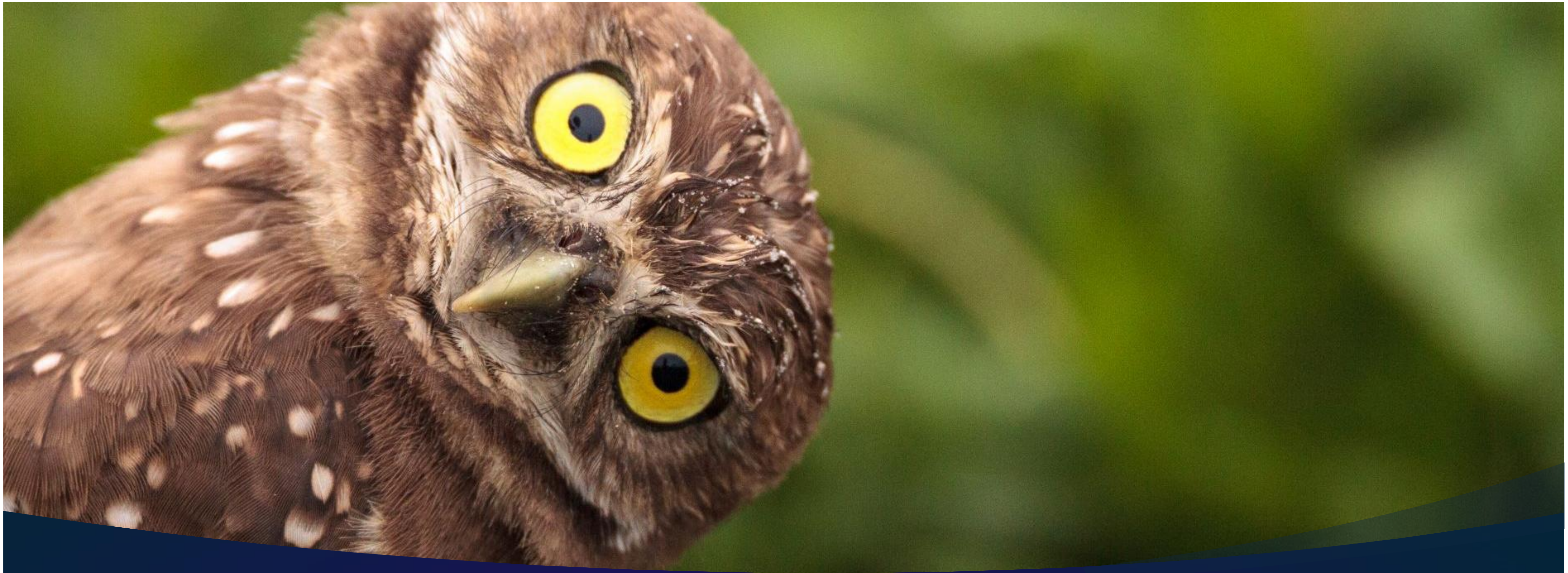
WizardLMTeam/WizardCoder-Python-34B-V1.0



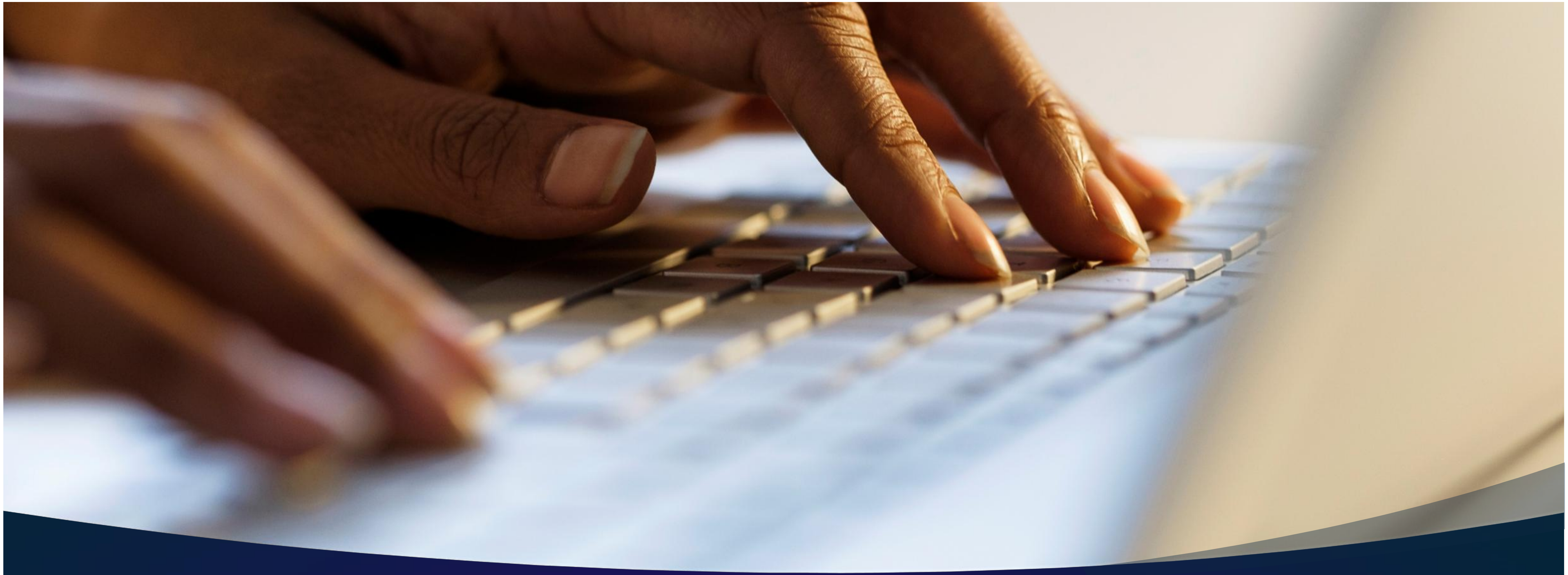
Demo Time

Key Takeaways

- Local LLMs are not difficult to install and customize
- Can be used by individuals and/or organizations for better privacy
- Customization increases relevance, ROI
- RAG lets you “interrogate” documents and data
- Knowledge is power – AI is here to stay, so learn as much as you can!



Q&A



Keeping Things Local

Making Your Own Private LLM

CORVUS | BRONWEN AKER

Corvus | Bronwen Aker
M.S. Cybersecurity, GSEC, GCIH, GCFE

- Website: `https://br0nw3n.com/`
- LinkedIn: `https://www.linkedin.com/in/bronwenaker/`
- Discord: `corvus_le_crow`

(Do your OSINT. I'm online.)