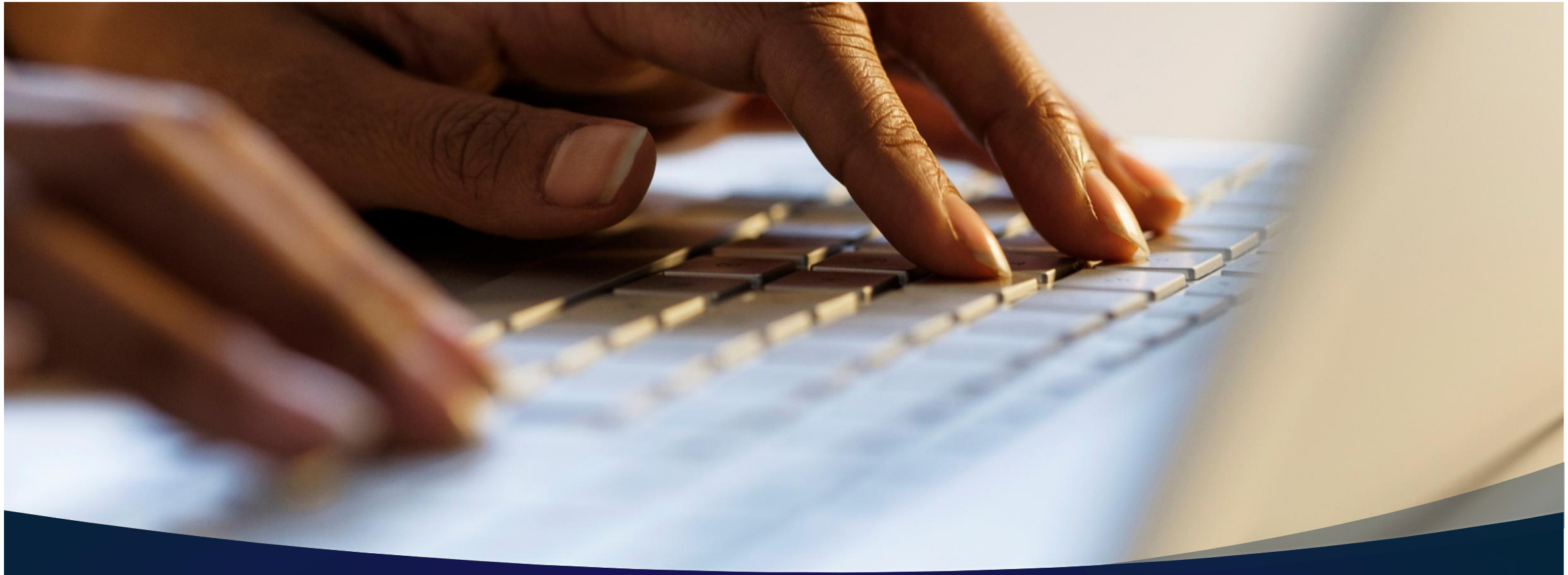


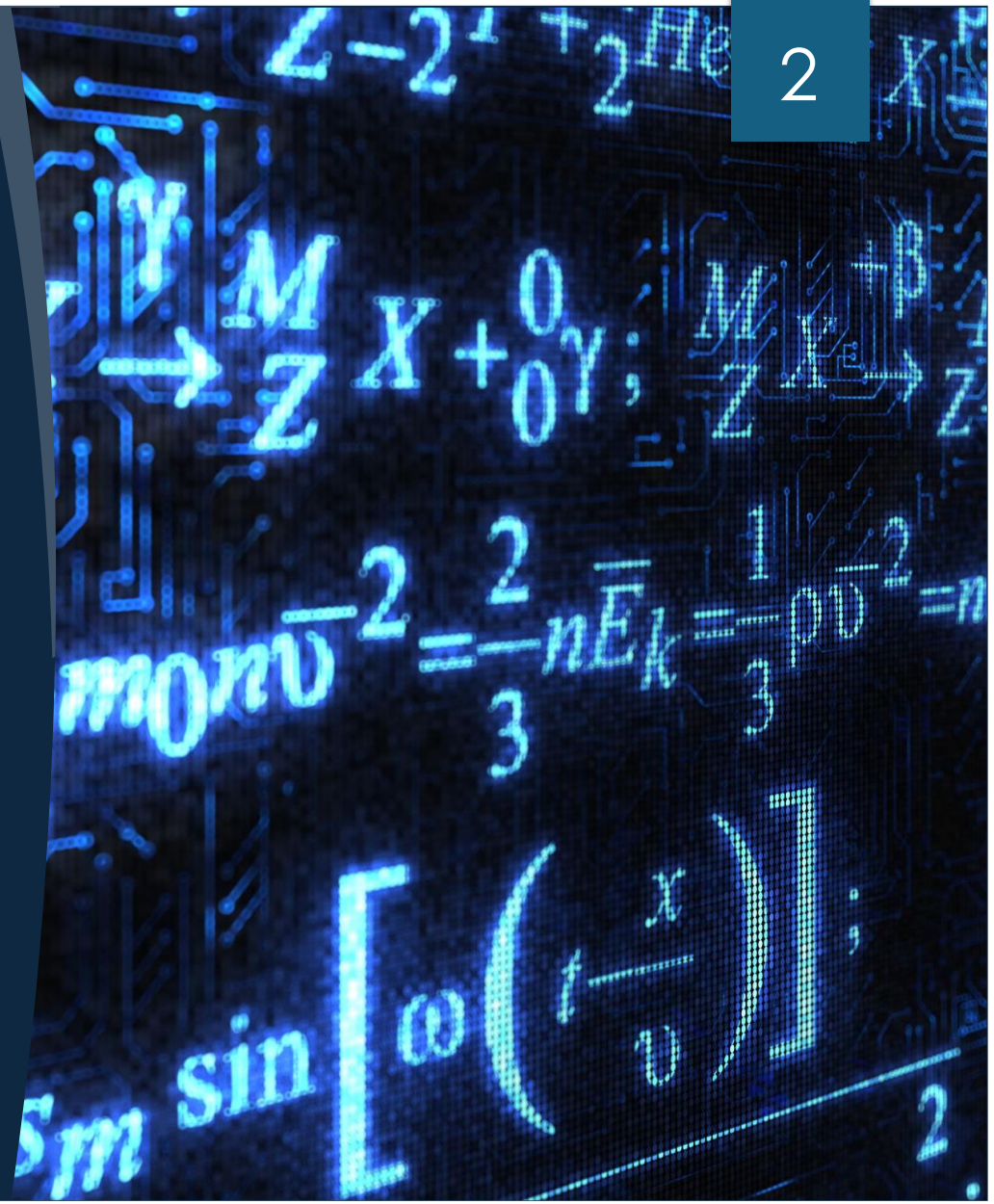
Keeping Things Local

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
  - Ollama, LM Studio, GPT4All, and more!
- Ways to Enhance Your LLM
- Demos:
  - Ollama Installation, 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 💩

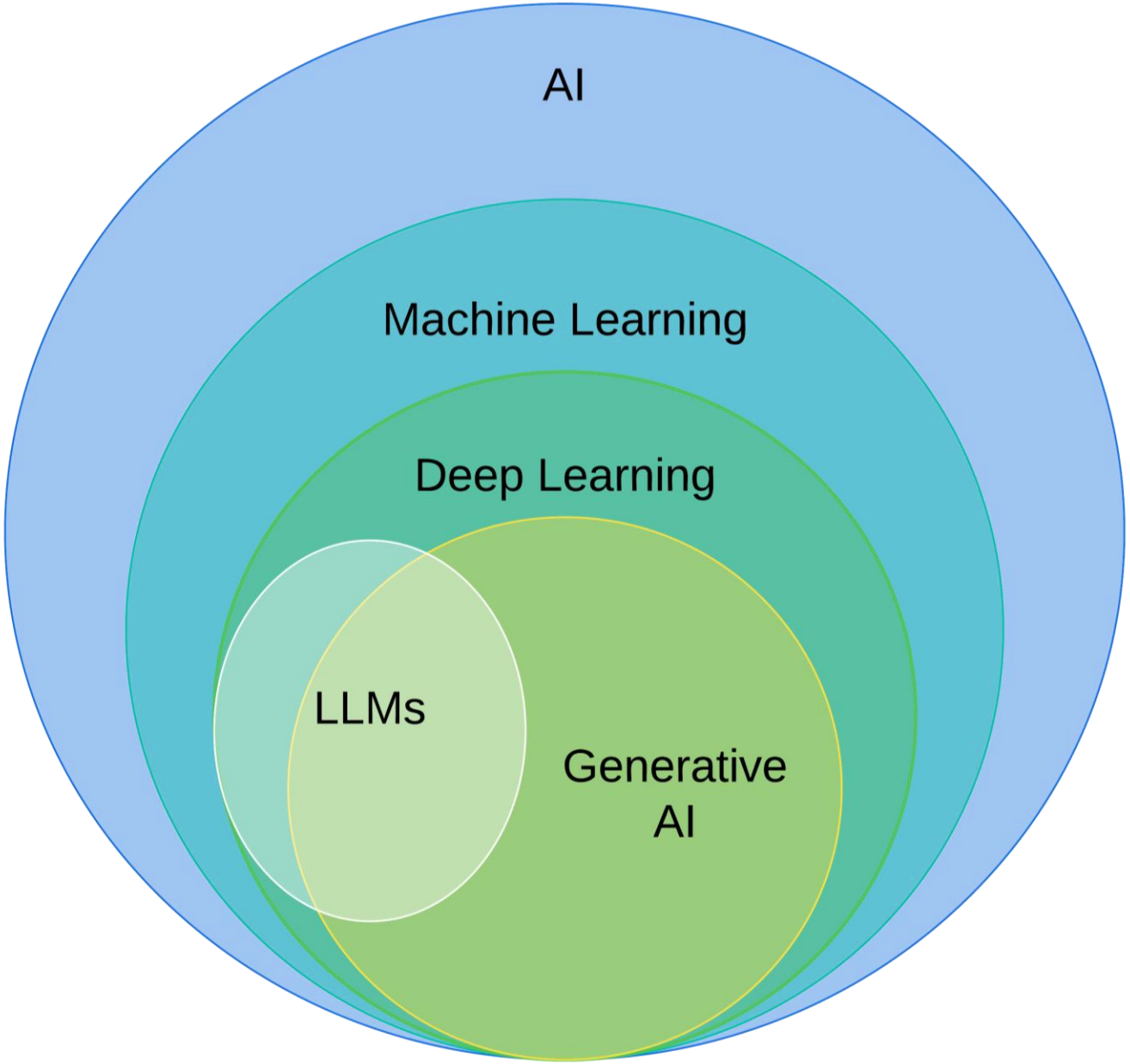
FULL DISCLOSURE

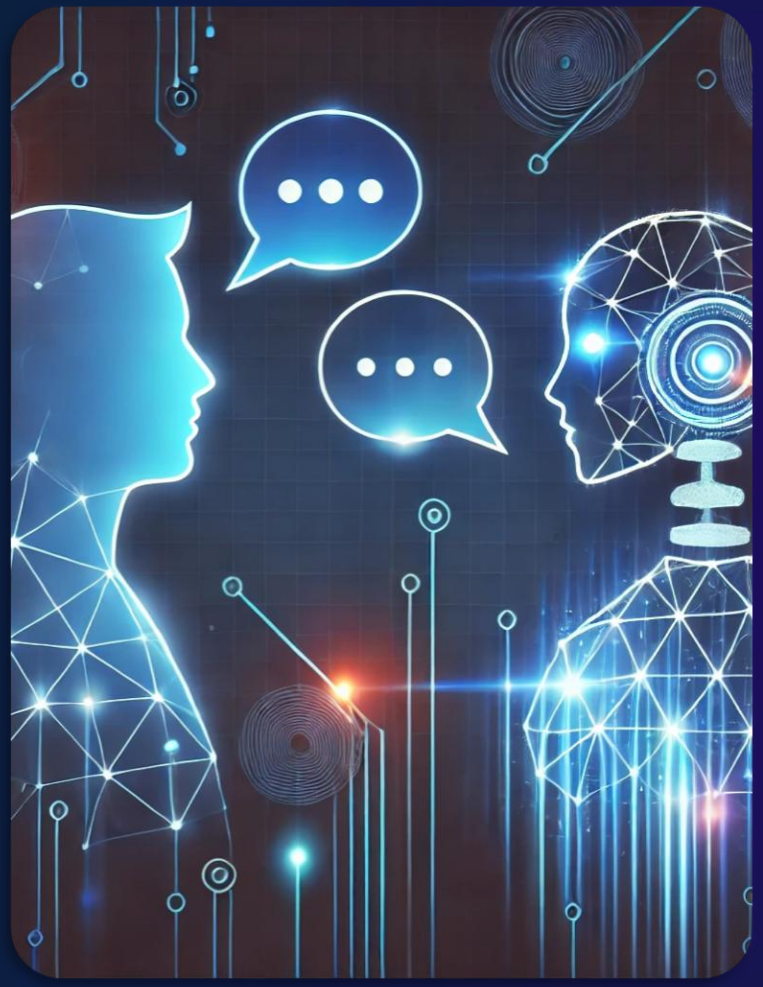
- This presentation was created using AIs like ChatGPT, Copilot, DALL-E, and Midjourney
- They were, at all times, under adult supervision*

* 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?

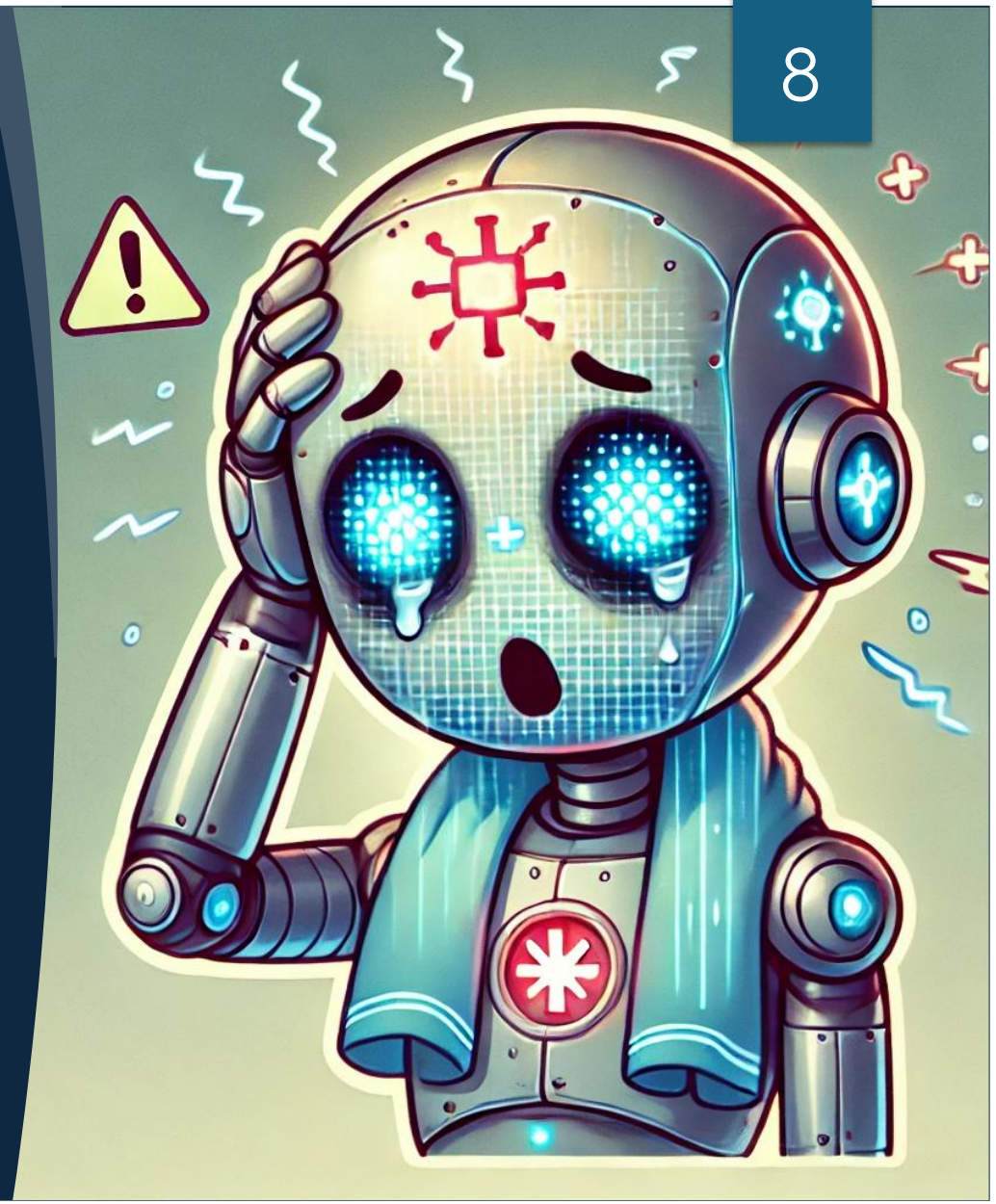
- LLMs are a subset of generative AI
- Text based
- Probabilistic
- Current models trained on MASSIVE amounts of data
- Most are general purpose
- Easy access to LLMs has popularized AI

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

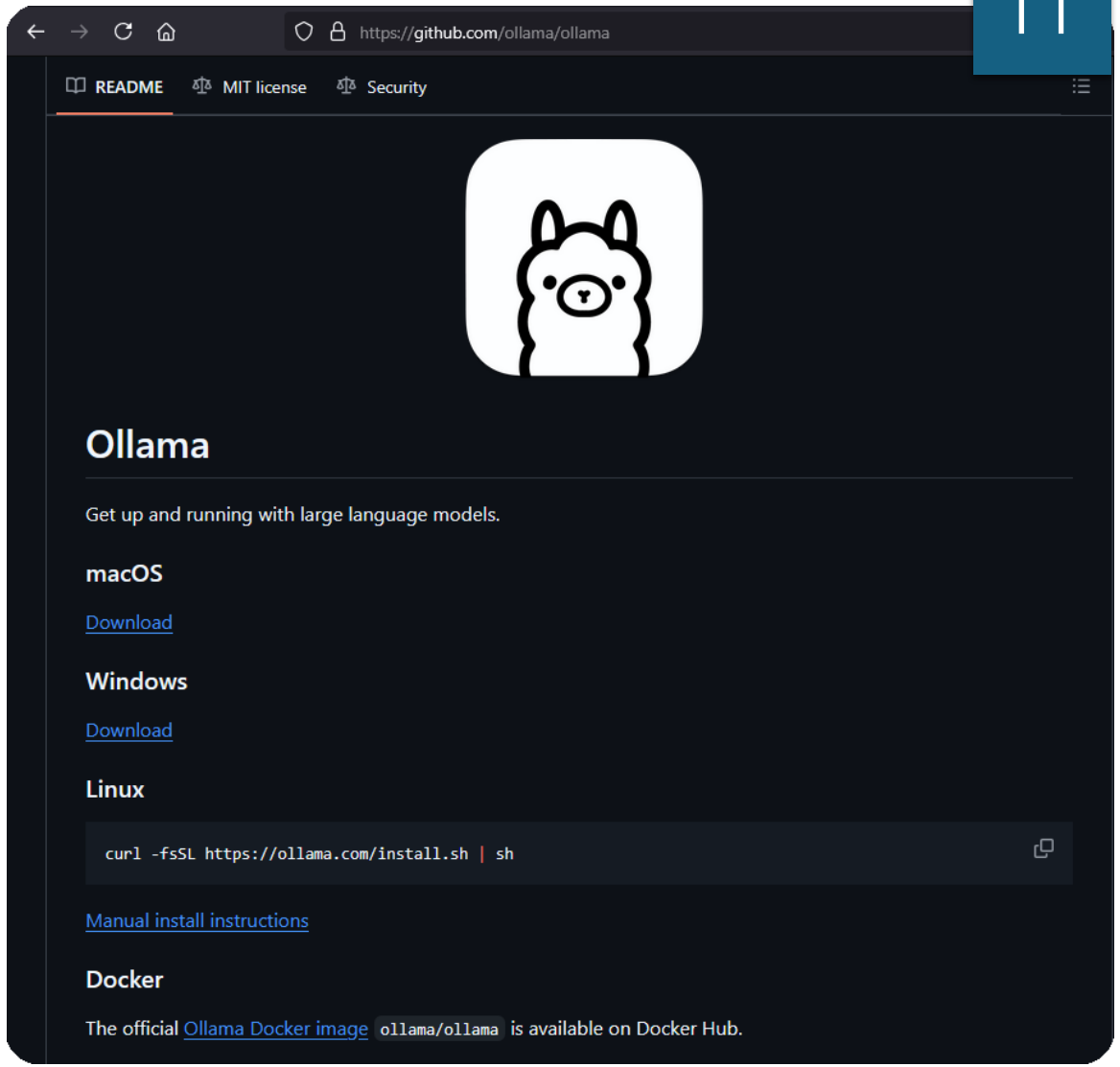
- Hardware
  - GPUs speed processing, but not 100% necessary
  - RAM & Disk Space: 16GB+ RAM, 50GB+ disk space
- Software
  - Determines UI, model management, other capabilities
  - Ollama, LM Studio, GPT4All
- LLM Model
  - Open-source models available from Hugging Face, Mistral, LLaMA, and others
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

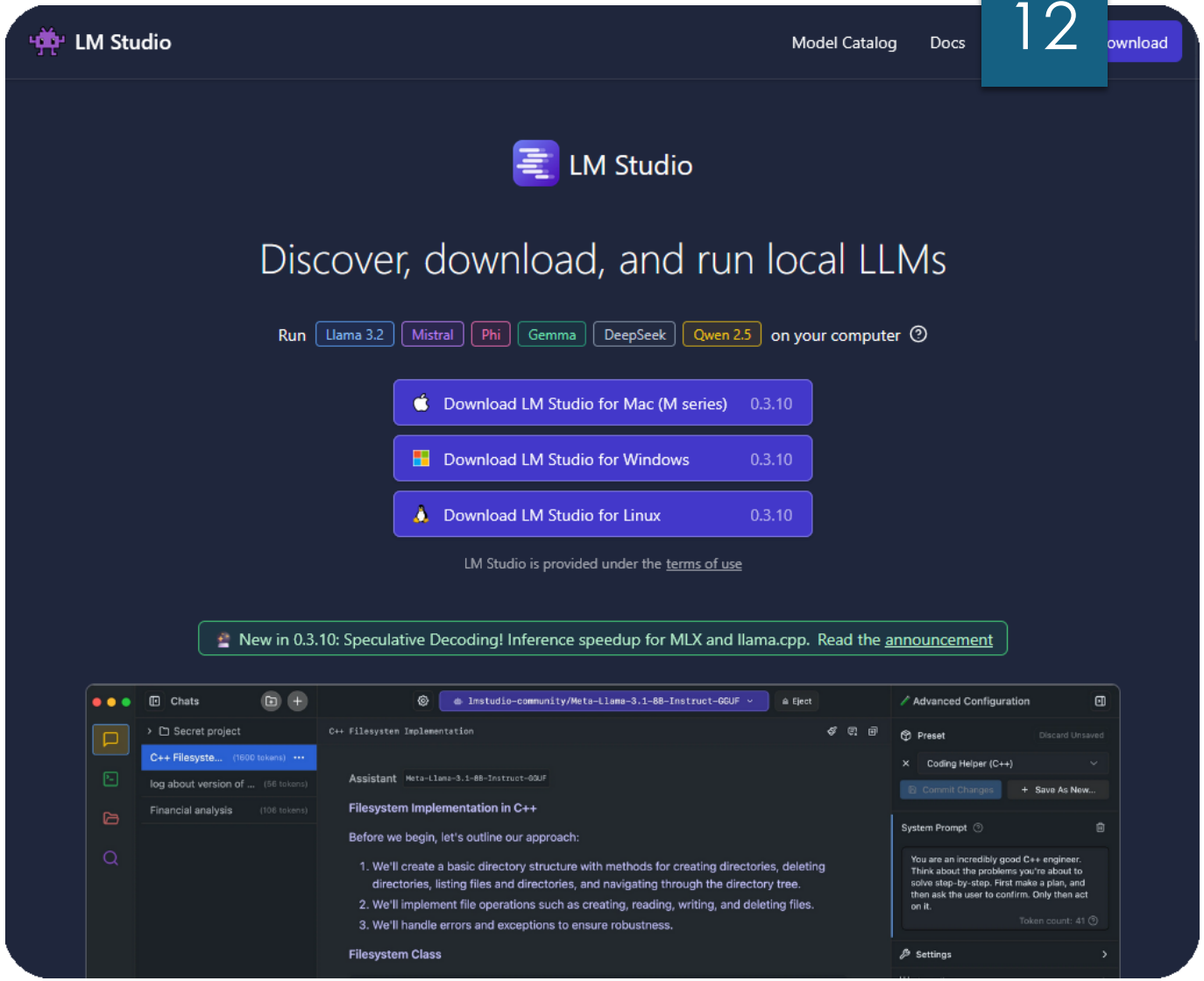
Software: Ollama

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



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



Many Other Software Options

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

LLM Model Repos

- Hugging Face – huggingface.co
  - Largest repository of open-source LLMs. Hosts models from Meta, Mistral, Falcon, and more.
- Ollama – ollama.com
  - Simple one-command installation for models like Mistral, Gemma, and Llama.
- Mistral AI – mistral.ai
  - Provides Mistral 7B, Mixtral, and other high-performance open models.
- Meta AI – ai.meta.com
  - Official source for LLaMA models (requires access approval).
- Google AI – ai.google.dev
  - Offers open and restricted-access models like Gemma and PaLM.

Choosing an LLM Model

- LLM Models != the same
  - Mistral 7B / Mixtral 8x7B
    Strong performance, open weights
  - LLaMA 2 (Meta)
    Optimized for reasoning; lacks fine-tuning
  - Falcon (Technology Innovation Institute)
    Competitive with GPT models
  - StableLM / OpenHermes
    Optimized for chat and instruction tasks
- Model Size
  - More parameters = more data/better accuracy, more compute needs
- General vs. Specialized
  - Models available for math, coding,

Ollama | GitHub - ollama/ollama: Get up | mistralai/Codestral-22B-v0.1 · H | Private browsing

https://ollama.com/

Discord GitHub Models

Search models

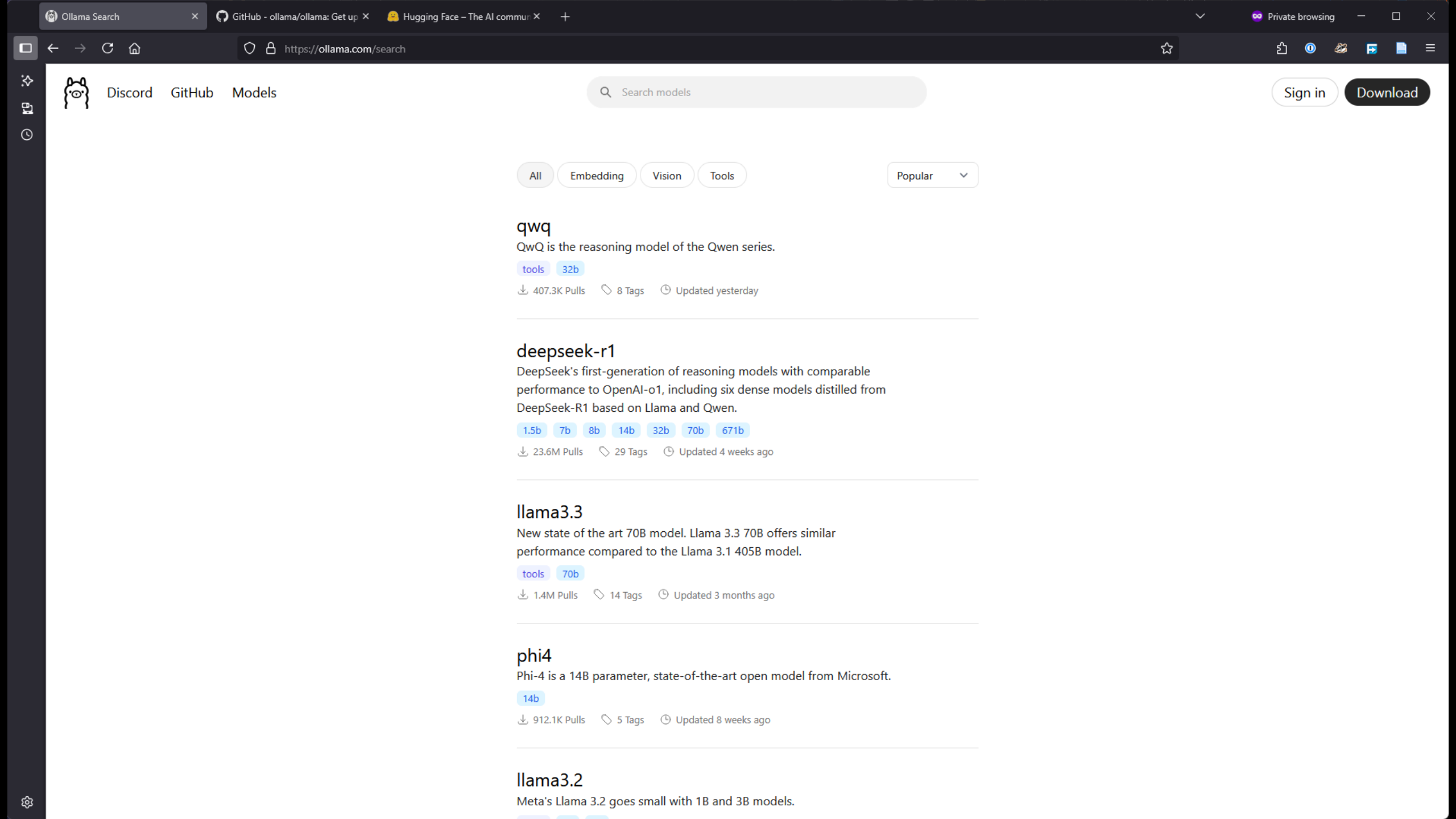
Sign in Download

Get up and running with large language models.

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

Download ↓

Available for macOS, Linux, and Windows


Ollama Search
GitHub - ollama/ollama: Get up
Hugging Face – The AI commun
Private browsing
https://ollama.com/search
Discord
GitHub
Models
Search models
Sign in
Download
All
Embedding
Vision
Tools
Popular
qwq
QwQ is the reasoning model of the Qwen series.
tools
32b
407.3K Pulls
8 Tags
Updated yesterday
deepseek-r1
DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.
1.5b
7b
8b
14b
32b
70b
671b
23.6M Pulls
29 Tags
Updated 4 weeks ago
llama3.3
New state of the art 70B model. Llama 3.3 70B offers similar performance compared to the Llama 3.1 405B model.
tools
70b
1.4M Pulls
14 Tags
Updated 3 months ago
phi4
Phi-4 is a 14B parameter, state-of-the-art open model from Microsoft.
14b
912.1K Pulls
5 Tags
Updated 8 weeks ago
llama3.2
Meta's Llama 3.2 goes small with 1B and 3B models.

code

All | Embedding | Vision | Tools | Popular

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

python

All | Embedding | Vision | Tools | Popular

python001/agent

pythondev/mymodel

this is my model

PythonProdigy/qwen2.5

xingyaow/codeact-agent-mistral

An LLM agent that deeply integrates with Python Interpreter.

1,089 Pulls | 1 Tag | Updated 11 months ago

me/llama3.1-python

843 Pulls | 1 Tag | Updated 7 months ago

ALIENTELLIGENCE/pythoncoderv2

AI Python Coder (An assistant to help with python coding tasks)

llama3

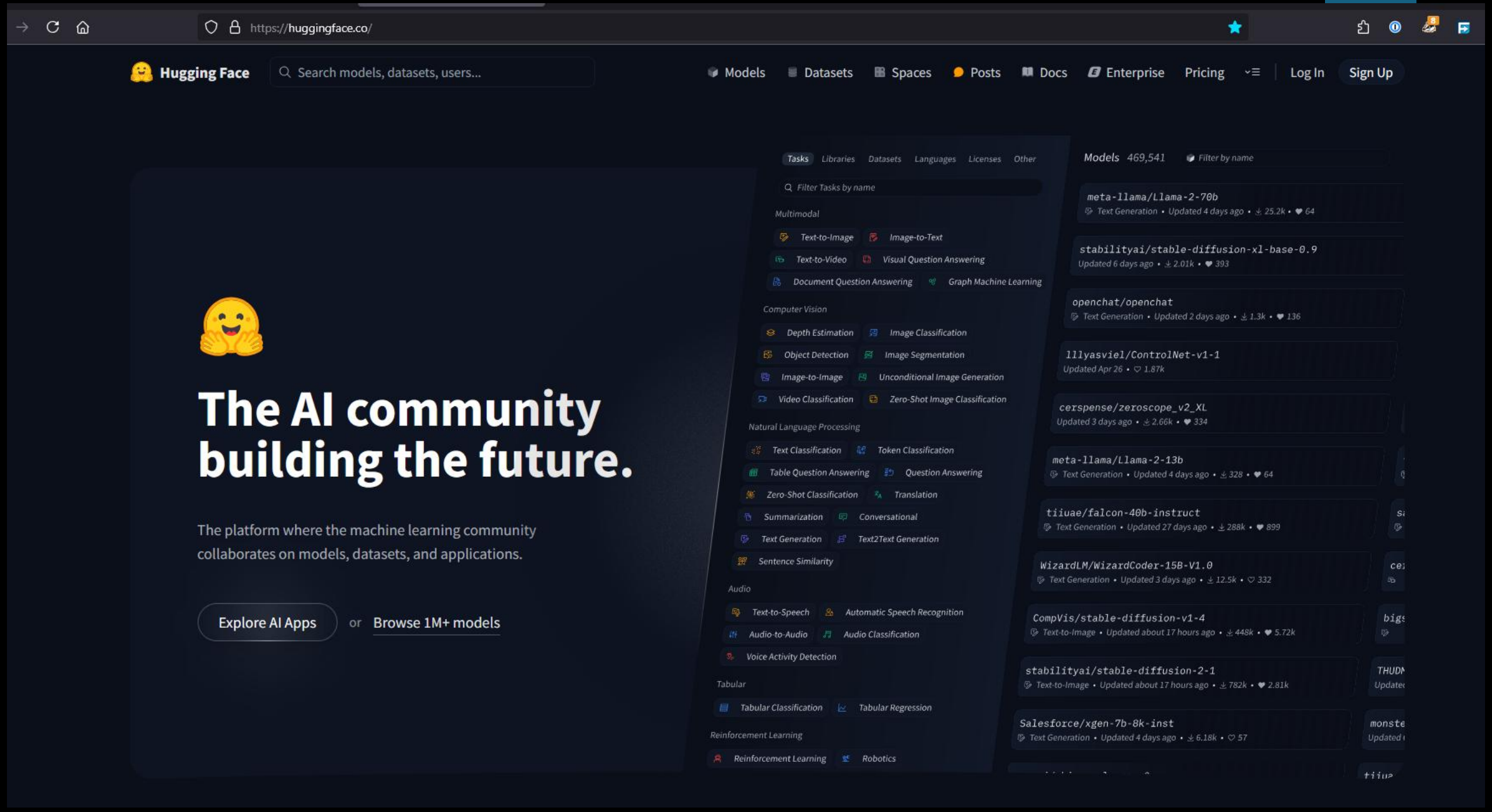
Meta Llama 3: The most capable openly available LLM to date

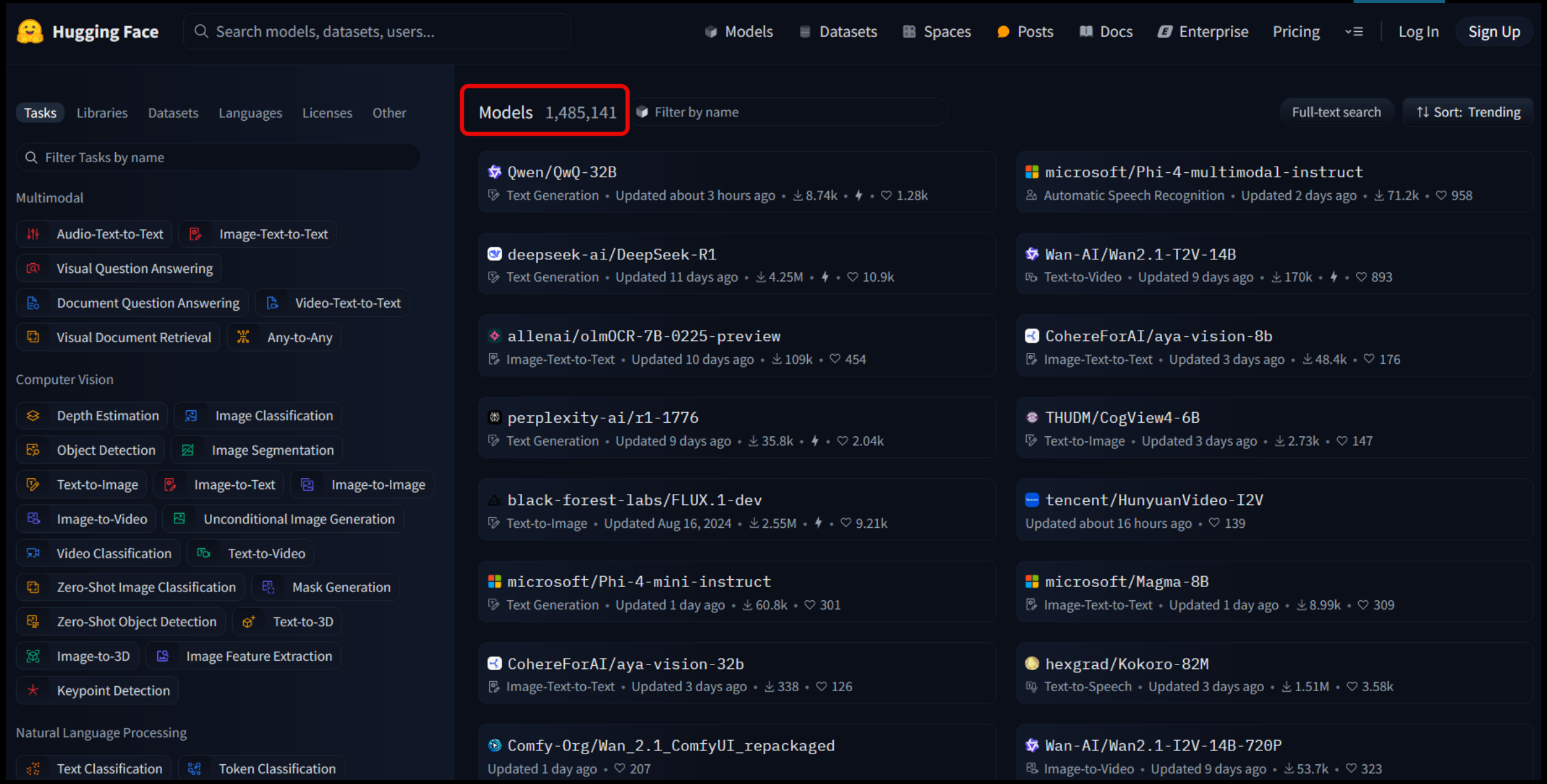
8b 70b

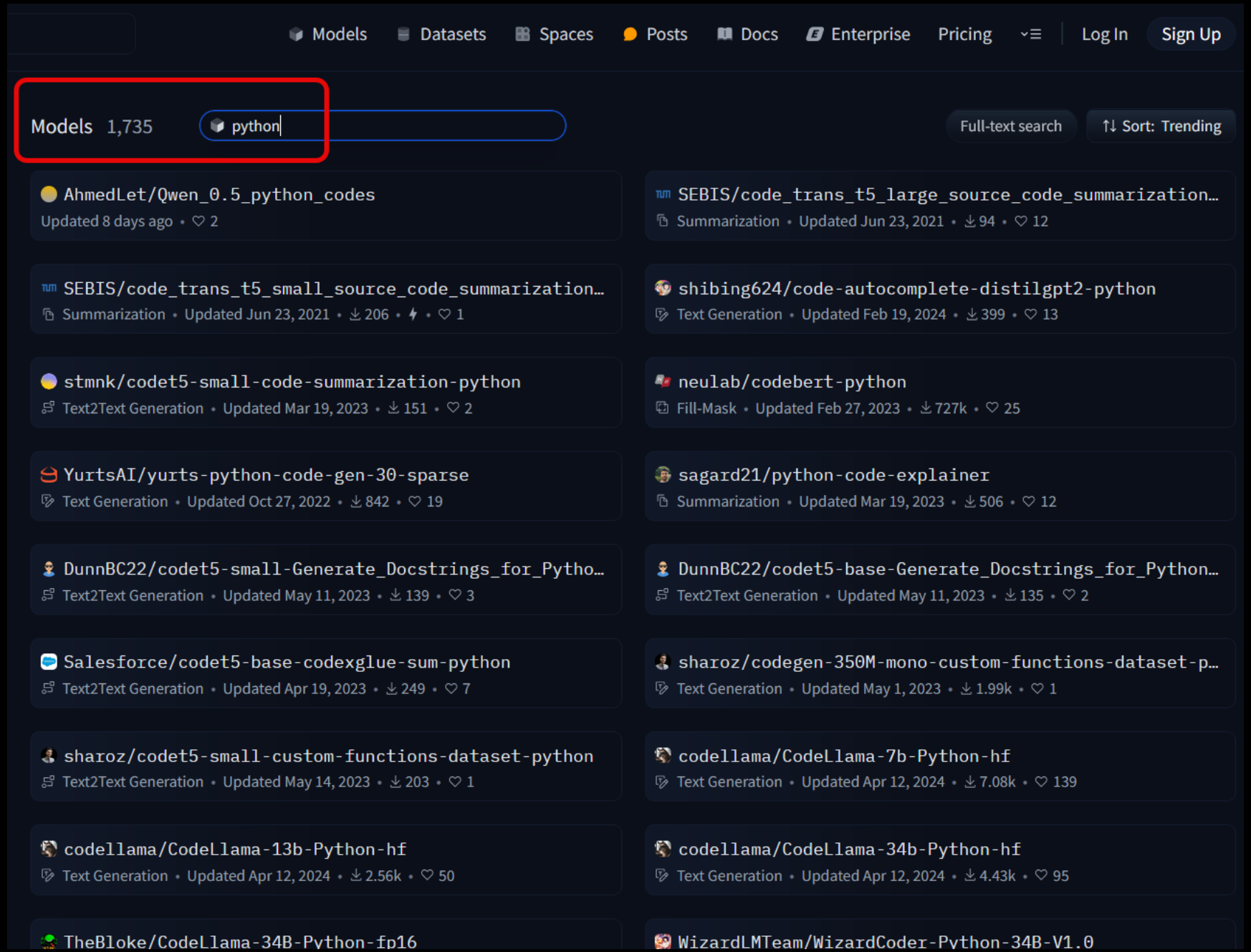
7.6M Pulls Updated 9 months ago

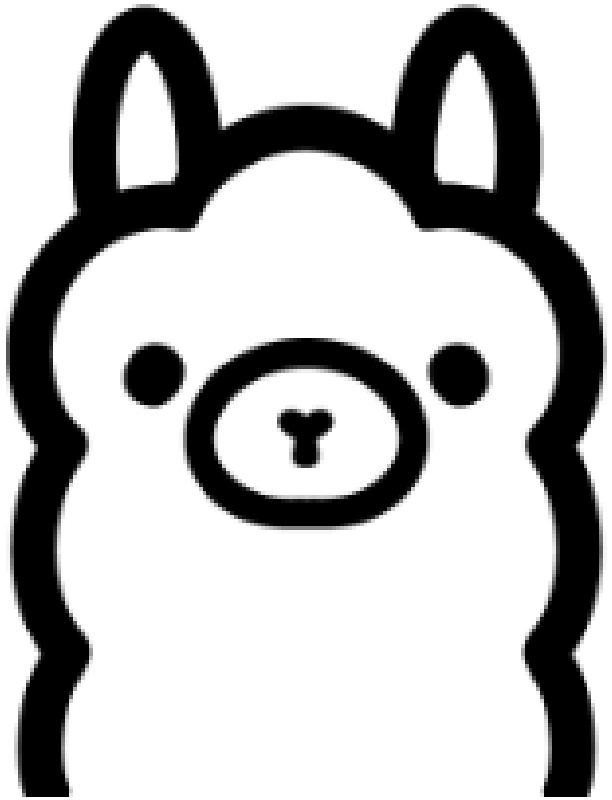
8b 68 Tags `ollama run llama3`

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


https://huggingface.co/
Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Posts
Docs
Enterprise
Pricing
Log In
Sign Up
The AI community building the future.
The platform where the machine learning community collaborates on models, datasets, and applications.
Explore AI Apps
or
Browse 1M+ models
Tasks
Libraries
Datasets
Languages
Licenses
Other
Filter Tasks by name
Multimodal
Text-to-Image
Image-to-Text
Text-to-Video
Visual Question Answering
Document Question Answering
Graph Machine Learning
Computer Vision
Depth Estimation
Image Classification
Object Detection
Image Segmentation
Image-to-Image
Unconditional Image Generation
Video Classification
Zero-Shot Image Classification
Natural Language Processing
Text Classification
Token Classification
Table Question Answering
Question Answering
Zero-Shot Classification
Translation
Summarization
Conversational
Text Generation
Text2Text Generation
Sentence Similarity
Audio
Text-to-Speech
Automatic Speech Recognition
Audio-to-Audio
Audio Classification
Voice Activity Detection
Tabular
Tabular Classification
Tabular Regression
Reinforcement Learning
Reinforcement Learning
Robotics
Models 469,541
Filter by name
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


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Posts
Docs
Enterprise
Pricing
Log In
Sign Up
Tasks
Libraries
Datasets
Languages
Licenses
Other
Filter Tasks by name
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
Models 1,485,141
Filter by name
Full-text search
Sort: Trending
Qwen/QwQ-32B
Text Generation • Updated about 3 hours ago • 8.74k • 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 • 10.9k
Wan-AI/Wan2.1-T2V-14B
Text-to-Video • Updated 9 days ago • 170k • 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 • 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 • 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

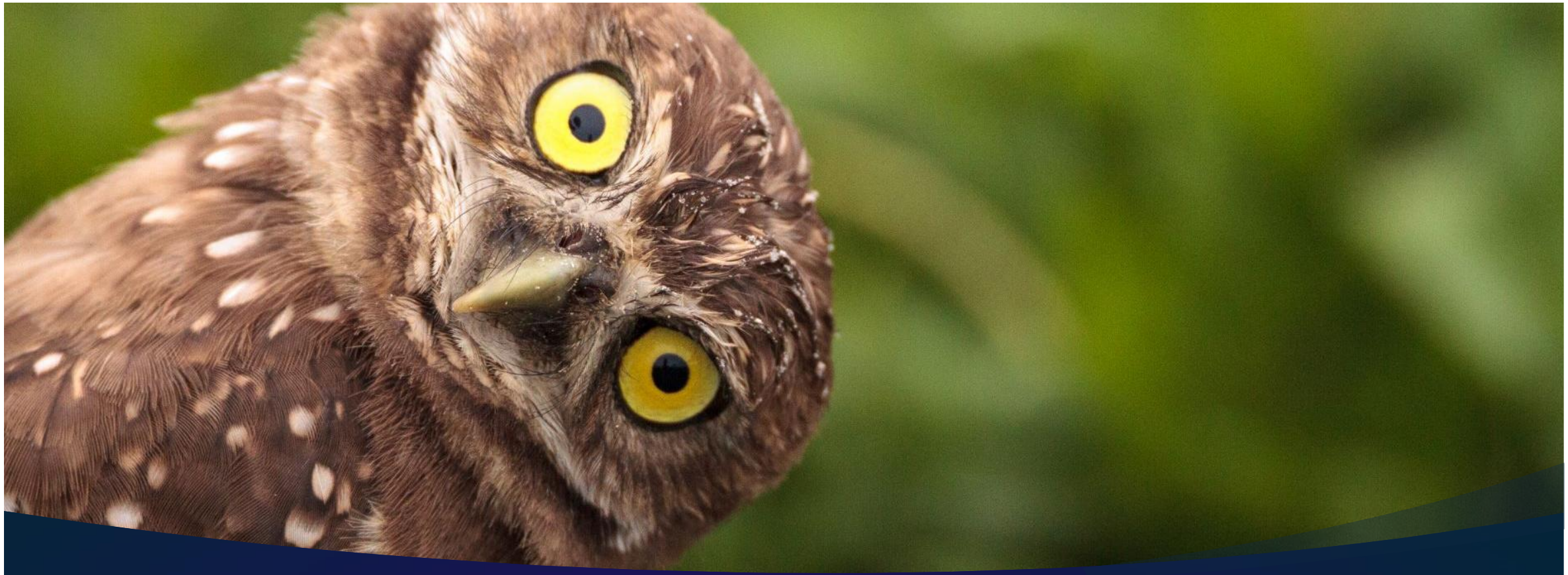




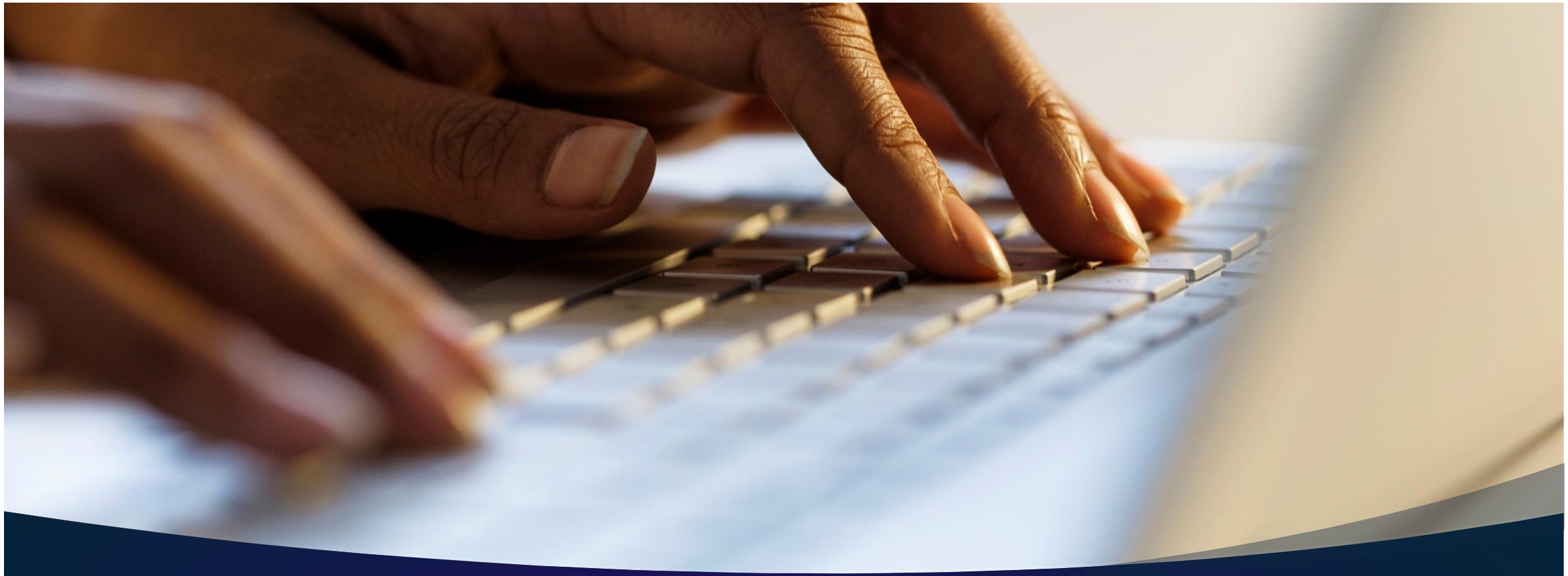
Demo Time: Ollama Installation

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



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.)