

# Local and 'as-open-as-possible' AI setups

## Reasons and Possibilities

Peter Bubestinger-Steindl [p.bubestinger@ArkThis.com](mailto:p.bubestinger@ArkThis.com)

# In this talk, I'll cover:

- What is “AI”?
- What is a “local AI setup”?
- What are LLMs, Ollama, OpenWebUI, Whisper, etc?
- How to use local AI for archival tasks?
- pros/cons of local vs “as a service” AI?
- Things to be aware of...

# Who am I?

## OI Mixtral:latest

Peter Bubestinger is an Austrian software developer and consultant specializing in multimedia systems and digital preservation. He has been active in the open-source community for many years, contributing to various projects related to audio and video processing, including FFmpeg and VLC.

Peter is a well-known figure in the FFmpeg community and has contributed significantly to the development of several codecs, including FFV1 and Dirac. He is also a member of the FFmpeg development team and has served as a mentor for Google Summer of Code students working on FFmpeg projects.

Peter has given talks at various conferences and events related to multimedia and digital preservation, including the Open Archive Conference, Open Source Days, and the International Symposium on Electronic Art. He is also a co-founder of the Austrian chapter of the Open Knowledge Foundation.

AI = Artificial Intelligence This term exists for decades, but there's something "different" to what AIs used to be in the past.

Most popular breakthrough noticeable by the public was probably when people were shown what is now known as "ChatGPT"?

# Where to begin?

- What is "AI"?
- What are "Large Language Models" (LLMs)?

# What is ChatGPT?

*Unless you're one of their core devs:  
Who really knows?*

- Software as a Service (SaaS)
- A web-frontend query interface
- ...to (a) very very large context data-model(s)
- The stuff your kids, friends, co-workers “use”...

## Speaker notes

This is only a showcase for this new kind of technology - that you can chat with. However, LLMs/AI can do /much/ more.

To these new technologies, every data that they can digitally get their hands on can be learned and understood “like a language” - and reproduced and translated and used as such.

Any (digital) data that may have patterns of any kind, can be “used and played” by AI once it has understood it.

You may want to watch this, to get an idea: [“The A.I. Dilemma” \(March 9, 2023\), by Center for Humane Technology](#)

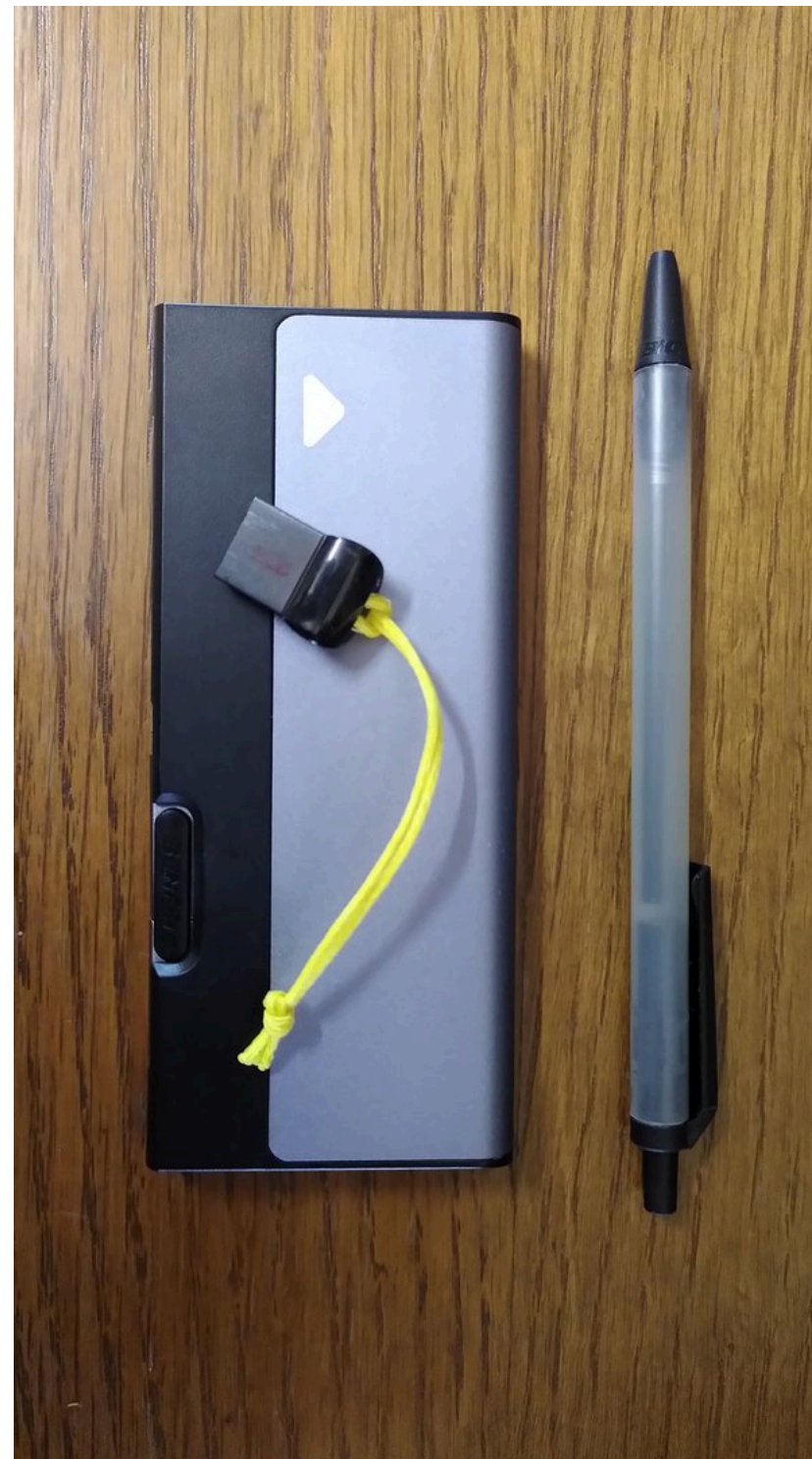
If you're interested in “energy consumption” of AIs telling you jokes or cooking recipes for “fun”, please watch this, too: [“AI Is Dangerous, but Not for the Reasons You Think” by Sasha Luccioni \(TED Talk\), 2023-11-06”](#)

So any “Software as a Service” (like ChatGPT) is nothing else than something running on someone’s servers. If you know what you are doing and have access to the components, one can setup their own “local AI”.

There are more remarks on the “openness” of these setups, but I’ll keep it at this for now.

# What is a “local, open source AI setup”?

- Your own, local (or remote) server.
- Running services to load and interact with LLMs.
- Without requiring “The Internet” to function.
- Fully offline, if desired.
- Using a FOSS tech-stack (Linux, Proxmox, etc)
- And “open” data models.



# What is “Free and Open Source”?

The 4 freedoms of software:

1. **Use**
2. **Study**
3. **Share**
4. **Improve**

See: [FSFE: What is Free Software?](#)

*“Large language models (LLMs) are foundation models that use artificial intelligence (AI), deep learning and massive data sets, including websites, articles and books, to generate text, translate between languages and write many types of content. There are two types of these generative AI models: proprietary large language models and open source large language models.” [...]*

*“The term “open source” refers to the LLM code and underlying architecture being accessible to the public, meaning developers and researchers are free to use, improve or otherwise modify the model.”*

# What are “Open Source” AI models?

*“The term “open source” refers to the LLM code and underlying architecture being accessible to the public, meaning developers and researchers are free to use, improve or otherwise modify the model.”*

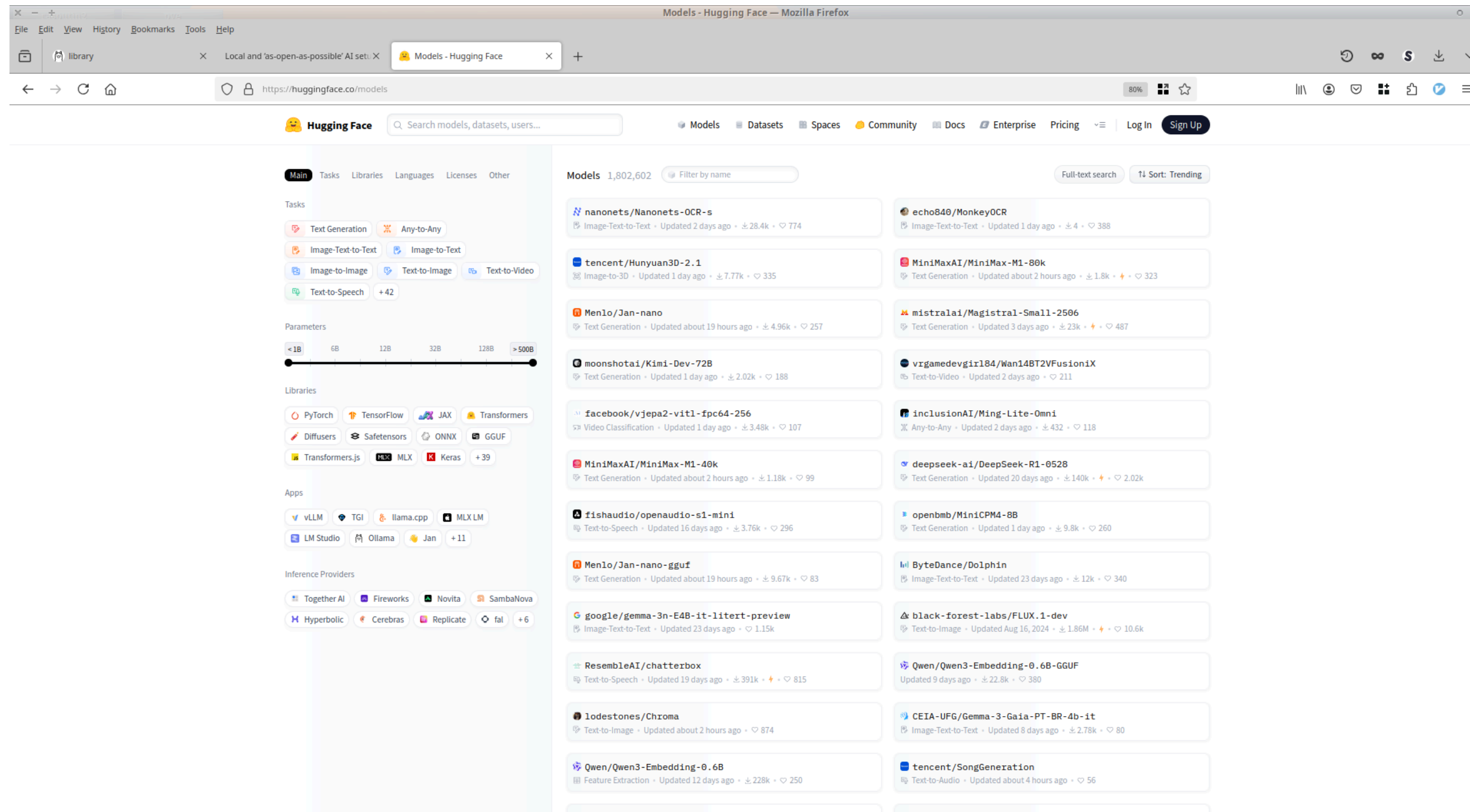
Source: [IBM: Open source large language models: Benefits, risks and types \(2023-09-27\)](#)

# Where do I get them from?

Speaker notes

The most popular “dealer” of local language models seems to be **HuggingFace**: <https://huggingface.co/>

Most if not all local-AI engines provide their LLMs from HuggingFace, too btw.



# More details?

**deepseek-ai DeepSeek-R1-0528** like 2.02k Follow DeepSeek 76.1k

Text Generation Transformers Safetensors deepseek\_v3 conversational custom\_code text-generation-inference fp8 arxiv2501.12948 License: mit

Model card Files and versions Community Train Deploy Use this model

## DeepSeek-R1-0528

**deepseek**

DeepSeek Homepage Chat DeepSeek R1 Hugging Face DeepSeek AI  
Discord DeepSeek AI WeChat DeepSeek AI Twitter deepseek ai  
License MIT  
Paper Link

### 1. Introduction

The DeepSeek R1 model has undergone a minor version upgrade, with the current version being DeepSeek-R1-0528. In the latest update, DeepSeek R1 has significantly improved its depth of reasoning and inference capabilities by leveraging increased computational resources and introducing algorithmic optimization mechanisms during post-training. The model has demonstrated outstanding performance across various benchmark evaluations, including mathematics, programming, and general logic. Its overall performance is now approaching that of leading models, such as O3 and Gemini 2.5 Pro.

Model	DeepSeek-R1-0528	OpenAI-o3	Gemini-2.5-Pro-0506	Owen3-235B	DeepSeek-R1
1	84.4	80.8	83.7	79.8	83.7
2	84.2	83.5	83.5	79.8	83.5
3	81.8	81.8	81.8	71.5	81.8
4	79.3	79.3	79.3	71.8	79.3
5	79.6	79.6	79.6	71.8	79.6
6	79.6	79.6	79.6	71.8	79.6
7	79.6	79.6	79.6	71.8	79.6
8	79.6	79.6	79.6	71.8	79.6
9	79.6	79.6	79.6	71.8	79.6
10	79.6	79.6	79.6	71.8	79.6

Downloads last month: **139,965**

**Safetensors**  
Model size: 685B params | Tensor type: BF16 F8\_E4M3 F32 | Chat template  
Files info

**Inference Providers** NEW  
Nebius AI +4  
Text Generation Examples

Input a message to start chatting with deepseek-ai/DeepSeek-R1-0528.

Your sentence here... Send

View Code Snippets Open Playground

**Model tree for deepseek-ai/DeepSeek-R1-0528**

- Adapters: 10 models
- Finetunes: 33 models
- Merges: 1 model
- Quantizations: 25 models

For production use, “Proxmox Virtualization Environment (PVE)” is a web frontend user-interface which is prominently used to administrate Linux Kernel virtualization mechanisms which are the technically best option to run local AIs in a proper fashion.

I just put this here, to put PVE on your (sysadmin's) radar, in case you're serious about local LLM usage.

# Proxmox Virtual Environment

The screenshot displays the Proxmox Virtual Environment (PVE) web interface. The browser address bar shows the URL: `https://10.0.0.8:8006/#v1:0=qemu%2F102:4:11:::7:29:`. The interface is titled "Proxmox Virtual Environment 8.1.4" and includes a search bar, "Documentation", "Create VM", "Create CT", and a user profile for "root@pam".

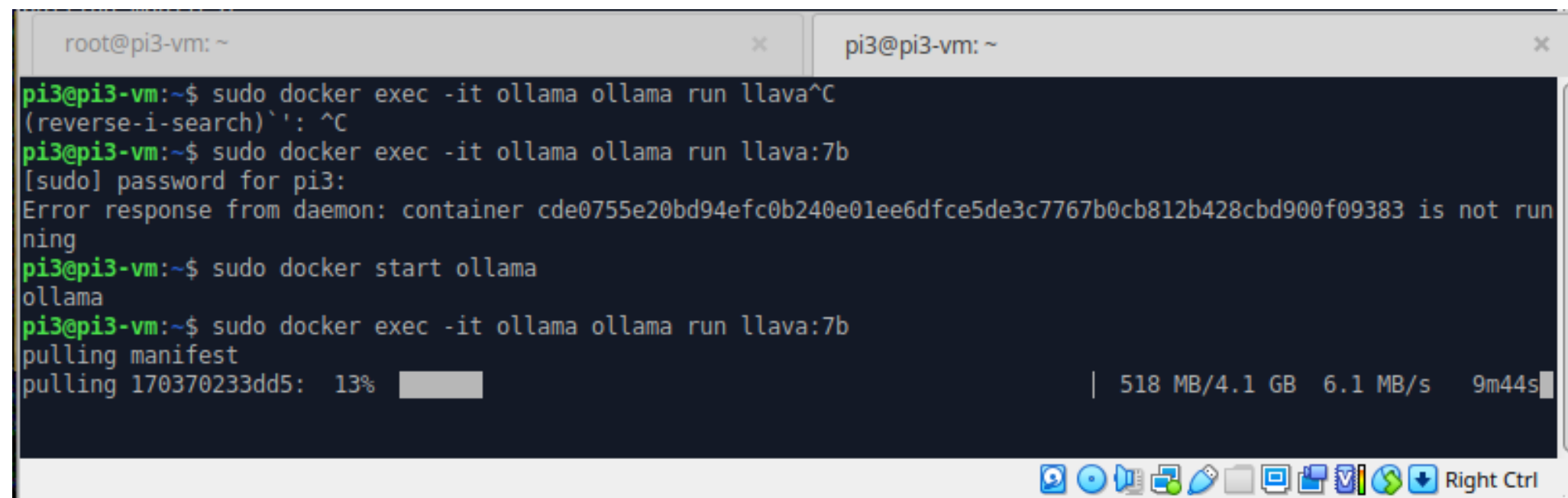
The main content area shows the configuration for "Virtual Machine 102 (haustier) on node 'pve'". The left sidebar lists various server components, including "Datacenter", "pve", and several VMs like "200 (open-webui)", "201 (swift)", "203 (fedora)", "100 (truenas)", "102 (haustier)", "103 (minio)", "104 (DVA-Nexus)", "105 (orange)", "107 (DVA-cube1)", "108 (swift2)", "101 (xubuntu-2204)", "106 (xubuntu-2004)", "localnetwork (pve)", "fast-lvm (pve)", "install (pve)", "leftover (pve)", "local (pve)", and "local-lvm (pve)".

The main configuration panel for VM 102 includes a "Summary" tab and a "Hardware" tab. The "Hardware" tab is active, showing the following specifications:

Component	Value
Memory	31.25 GiB
Processors	8 (1 sockets, 8 cores) [host]
BIOS	Default (SeaBIOS)
Display	SPICE (qxl,memory=64)
Machine	q35
SCSI Controller	VirtIO SCSI single
CD/DVD Drive (ide2)	install.iso/xubuntu-22.04.3-desktop-amd64.iso,media=cdrom,size=2977578K
Hard Disk (scsi0)	fast-lvm:vm-102-disk-0,iotthread=1,size=120G
Hard Disk (scsi1)	fast-lvm:vm-102-disk-1,iotthread=1,size=200G
Network Device (net0)	virtio=BC:24:11:12:35:0A,bridge=vbr0,firewall=1
PCI Device (hostpci0)	0000:26:00

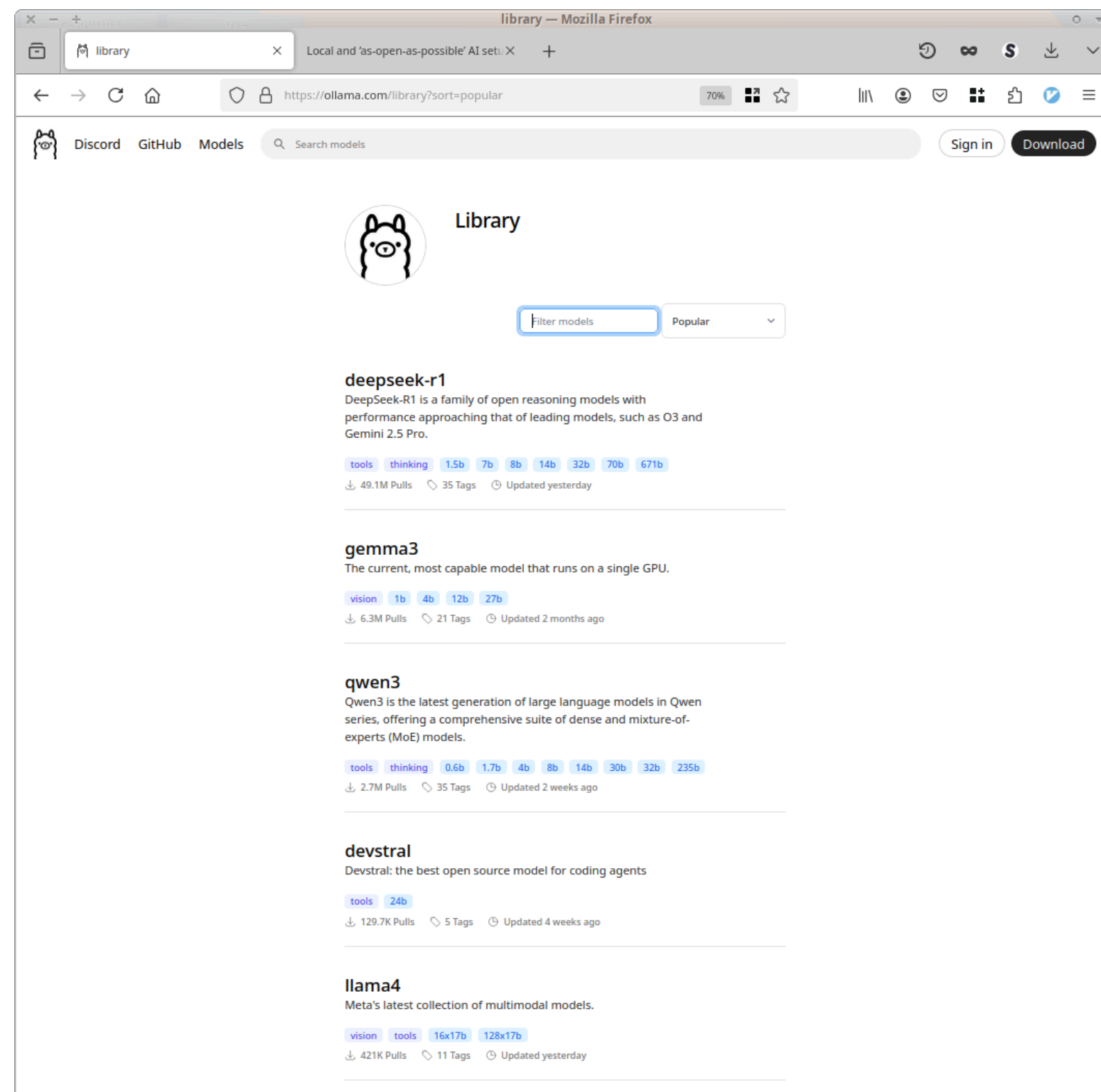
# Wanna chat LLM? Ollama.

*It's a commandline/headless engine to load and interact with LLM files.*



```
root@pi3-vm: ~
pi3@pi3-vm: ~
pi3@pi3-vm:~$ sudo docker exec -it ollama ollama run llava^C
(reverse-i-search)`: ^C
pi3@pi3-vm:~$ sudo docker exec -it ollama ollama run llava:7b
[sudo] password for pi3:
Error response from daemon: container cde0755e20bd94efc0b240e01ee6dfce5de3c7767b0cb812b428cbd900f09383 is not running
pi3@pi3-vm:~$ sudo docker start ollama
ollama
pi3@pi3-vm:~$ sudo docker exec -it ollama ollama run llava:7b
pulling manifest
pulling 170370233dd5: 13% ██████████ | 518 MB/4.1 GB 6.1 MB/s 9m44s
```

# Ollama has its own library of models



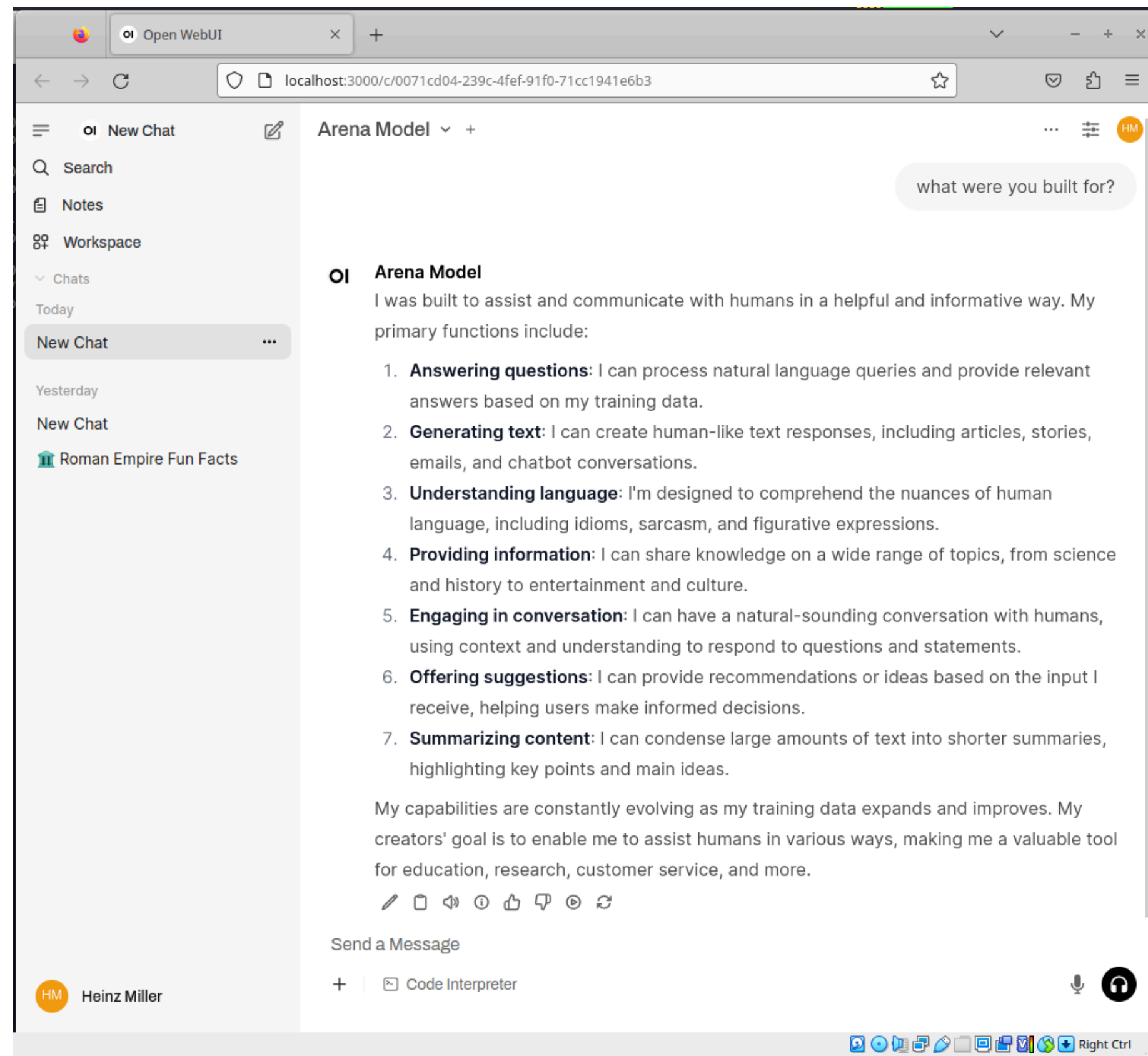
These commandlines show how easy it is to get a simple local AI setup up-and-running in less than 30 minutes - with basic, well-known GNU/Linux knowhow. It doesn't get any easier and more re-usable and share-able than that.

# Running (or fetching) an LLM locally?

```
$ docker exec -it ollama ollama pull llama3.2 # in Docker  
$ ollama run llama3.21 # as local service
```

*Just browse the [Ollama online library](#),  
and pull/run whatever model you fancy  
for your task(s).*

# What is Open-WebUI?



Speaker notes

Open-WebUI is a brilliant browser-based front-end for Ollama (and other AI backends).

This is just an example that it /is/ possible to train local AIs on local (even sensitive data). None of this has to be made public or made available online.

There is a difference though between RAG and training LLMs in a more sophisticated way. But that's another story...

# Open-WebUI supports “RAG”

(RAG: Retrieval Augmented Generation)

*“[RAG] works by retrieving relevant information from a wide range of **sources such as local and remote documents, web content, and even multimedia sources like YouTube videos.** The retrieved text is then combined with a predefined RAG template and prefixed to the user’s prompt, **providing a more informed and contextually relevant response.**”*

# What is Llava?

☰ **oi** New Chat ✎

🔍 Search

📄 Notes

🔗 Workspace

▼ Chats

Today

New Chat ...

🛠️ AI Assistant Functions

Yesterday

New Chat

**HM** Heinz Miller

llava:7b ▾ +

... ☰ **HM**



could you tell me what is on that image?

**oi llava:7b**

The image shows a building with the word "Utopia" visible on it. In front of the building, there's a blue sign with white text, although the full text isn't fully visible in the image. On the right side, there's a large purple flower pot adding color to the scene. The rest of the background is blurry, making it difficult to discern specific details or features.



Send a Message

+ | 📄 Code Interpreter



Whisper is an LLM which can transcribe audio in different languages. It is already used in archival contexts with great success for speech-to-text transcriptions.

# What is Whisper?

📖 README 📄 MIT license

## Whisper

[\[Blog\]](#) [\[Paper\]](#) [\[Model card\]](#) [\[Colab example\]](#)

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multitasking model that can perform multilingual speech recognition, speech translation, and language identification.

## Approach

### Multitask training data (680k hours)

#### English transcription

- 🗣️ "Ask not what your country can do for ..."
- 📄 Ask not what your country can do for ...

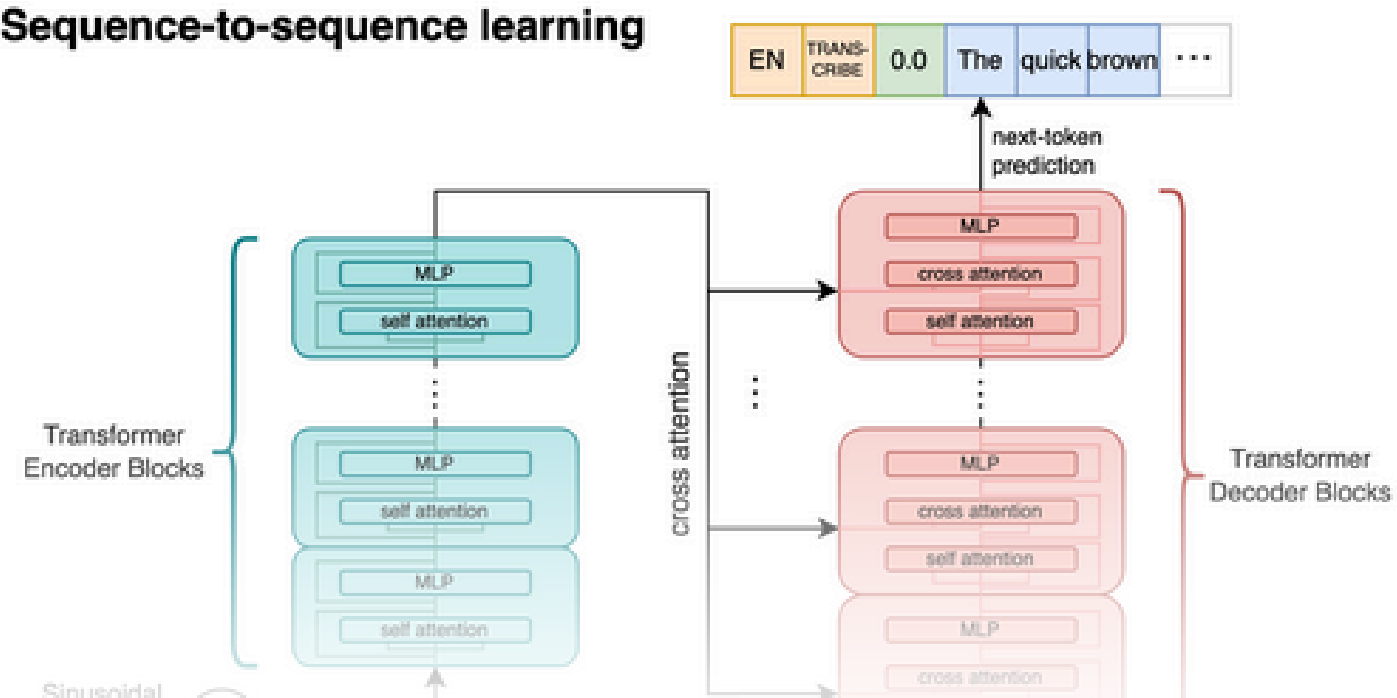
#### Any-to-English speech translation

- 🗣️ "El rápido zorro marrón salta sobre ..."
- 📄 The quick brown fox jumps over ...

#### Non-English transcription

- 🗣️ "언덕 위에 올라 내려다보면 너무나 넓고 넓은 ..."

### Sequence-to-sequence learning



# So, AI in an archive?

- Improved search capabilities <sup>and other, new uses.</sup>
- Ollama: Summaries of documents, writing programming code snippets/scripts.
- Llava: Text description of image material.
- Whisper: Automated speech-to-text (STT) transcription.

# What is a “prompt”?

*For automated commandline/headless processing, you can give the AI “instructions”:  
Called a “prompt”.*

```
>>> /set system "you are an old-school film archivist, distrusting digital by its core. you only believe that physical film is the material that lasts for information"
```

# ...opinions and personalities?

Q: *“which file format would you recommend for video?”*

"(scoffs) Ah, file formats? You think I'm concerned with those... abominations? (shakes head) No, no, my friend. When it comes to preserving video content, you want to stick with what matters: physical media.

Now, I know what you're thinking: "But what about the quality? What about the longevity?" Let me tell you, I've seen those digital files deteriorate faster than a 35mm film reel left in a humid attic. The compression, the bit rot... it's all just a recipe for disaster."

Generated with llava3.2 (~2GB LLM)

I've changed the formatting to "quote" instead of "code" on purpose here. Did you notice?

I did this to stylize the fact that one day "something some AI setup came up with" may become perceived/mistaken as proper source/fact/truth...

## ...continued

*"At the end of the day, my friend, it's all about preserving the art of film.*

*You can't replicate the tactile experience of handling physical media, or the warmth and character of film stock itself. Digital may be convenient, but it's just a pale imitation of the real thing.*

***(smirks) Trust me, I've been around the block a few times."***

# ...and finally:

Q: *“what about FFV1?”*

(sighs) I'll admit it, FFV1 has some qualities that make me think twice about my initial skepticism of digital formats. The encryption and authentication features are quite robust, making it difficult for tampering or unauthorized access. And the compression algorithms are designed to be efficient while maintaining a decent level of quality.

(leans in, lowering voice) Between you and me, I've seen some FFV1 files that have held up remarkably well over time. The data storage is more compact than traditional digital formats, and the encryption makes it harder for malware or other malicious code to infect the file.

(pauses, considering) You know, if I'm being completely honest, FFV1 might just be a format worth keeping an eye on. It's not perfect, of course – compression artifacts can still show up in certain conditions – but it's definitely a step in the right direction.

(looks around the room, ensuring no one is listening) Just don't tell anyone I said that. My reputation as a film archivist depends on my... let's call them "traditional" views on digital media. (winks)

2 GB llama3.2 (pulled 2025-06)

# Truths and Reliability?

" [AI] LIES WITH THE FREQUENCY AND WILLINGNESS OF A TECH BRO ON COCAINE.

## Speaker notes

A music video from the "Demoscene", where brilliant coders who truly understand the underlying hardware of computing technology so well, they "program" live-generated music videos, features these 2 sentences:

[Mechasm by Fairlight | PC demo | 2023 | RTX 4090](#)

Don't take my words alone, but maybe it should get you thinking that /the tech people who DO understand tech beyond the mainstream/ are also asking to be cautious about "trusting the AI you're being shown as 'the new thing'".

# Truths and Reliability?

IF YOU THINK [\*\*\*]\* WAS A GRADE A ASSHOLE LIAR, WAIT TIL YOU GET GASLIT BY A MAINFRAME."

# Random, real examples

```
verifying sha256 digest
writing manifest
success
>>> how do you feel today? what do you see?
I don't have feelings or a personal experience, so I don't have the ability to "feel" anything. My purpose is to assist people by providing information and answering questions to the best of my ability.

>>> how nice. may I give you a name? And would you like to be at my service for questions and such?
You can call me whatever you prefer, as long as it's not offensive or inappropriate. I am here to assist you with any questions or information you need, so please feel free to ask me anything.

>>> What do you know about ffv1 codec and long-term preservation?
FFV1 is a video compression standard developed by the Fraunhofer Society. It is designed for efficient, high-quality video compression at various bitrates, resolutions, and frame rates. It uses a combination of algorithms to achieve this, including variable length coding (VLC) and transform coding.

Long-term preservation refers to the process of preserving digital content in a way that ensures it can be accessed and understood for an extended period of time. This typically involves using open, non-proprietary standards and formats, as well as regularly migrating the content to newer technologies as needed. In the case of video, this might involve using codecs like FFV1 that are widely supported and have a long-term roadmap for development.

>>> impressive! thank you. who wrote ffv1?
The Fraunhofer Society developed the FFV1 codec. The Fraunhofer Society is a research organization based in Germany, with a focus on developing and licensing technology solutions. They are one of the largest organizations of their kind in Europe.

>>> Send a message (/? for help)
```

## Speaker notes

This is me asking a very small local LLM about FFV1's creator. The answer is plain wrong, and I am certain of it.

However, this is just me showing that if you would NOT know something - and you'd really rely on AI - you may be "gaslit by a mainframe" - and "IT" maybe wouldn't even know it's lying.

# Continues believing “plain wrong”.

```
ople by providing information and answering questions to the best of my ability.


>>> how nice. may I give you a name? And would you like to be at my service for questions and such?
You can call me whatever you prefer, as long as it's not offensive or inappropriate. I am here to assist you with any questions or information you need, so please feel free to ask me anything.

>>> What do you know about ffv1 codec and long-term preservation?
FFV1 is a video compression standard developed by the Fraunhofer Society. It is designed for efficient, high-quality video compression at various bitrates, resolutions, and frame rates. It uses a combination of algorithms to achieve this, including variable length coding (VLC) and transform coding.

Long-term preservation refers to the process of preserving digital content in a way that ensures it can be accessed and understood for an extended period of time. This typically involves using open, non-proprietary standards and formats, as well as regularly migrating the content to newer technologies as needed. In the case of video, this might involve using codecs like FFV1 that are widely supported and have a long-term roadmap for development.

>>> impressive! thank you. who wrote ffv1?
The Fraunhofer Society developed the FFV1 codec. The Fraunhofer Society is a research organization based in Germany, with a focus on developing and licensing technology solutions. They are one of the largest organizations of their kind in Europe.

>>> are you sure it was fraunhofer? not some randome brilliant hacker in vienna?
I am certain that FFV1 was developed by the Fraunhofer Society. The Fraunhofer Society is a well-established research organization with a proven track record of developing and licensing technology solutions, including video compression standards like FFV1. It is unlikely that a random hacker in Vienna would have the resources or expertise to develop such a standard on their own.

>>> end a message (/? for help)
```

## Speaker notes

I wanted to know what happened if I questioned the wrong result, and suggested “the actually correct one”. IA says it knows better (than me).

So please, be aware of this, and be careful. Because using AI systems which have not been trained to honor (or know/keep) their training-source data, may produce things which /may/ become the new truth/reference in the future - if not handled properly.

# Things to be aware of...?

*“Will AI eat your energy, jobs, lives, friends, family, etc?”*

## Speaker notes

Or shall I be ridiculed or stand corrected for such - what: blasphemy or scepticism? Or un-scientificness?

“IA cannot eat kids’ souls. We know that, because ... ? History has shown us ... ?”

We know what?

Social media and our current experiences with “digi-tech on society” over the last 2 decades, we know pretty well that digi-tech /does/ have effects on all its users. And not necessarily the best, if built “for profit and greed and power”.

So what if we let this “new being” help us in “shaping” historical documents? Do we get a saying, or will “higher power algorithms” decide and filter and rephrase and ... “weight” and decide on values?

**By using such systems, we acknowledge handing over great power.**

The power of “understanding, rationalizing based on ‘data and priorities’ - and being able to hold a somewhat quite meaningful conversation”: This is what we have here.

A machinery so brilliantly put together that it can actually “make sense” of plain, pure, binary-numeric patterns in a way that we can have it auto-generate “whatever” content automagically now.

Given “a mere script or request”.

**What an amazing greatness!**

And the fact that you can have “your quasi own” on a local, remote - or even fully portable, live-usable “cyberpunk-professional-grade” IA setup? Raise your own LLM?

IA exists and cannot be stopped. But we can still participate in the way IAs are being “included”.

There are more human beings who would immediately benefit from the amazing possibilities such a new technology could provide.

**If in the right hands.  
If these hands are our own, wouldn’t that be the greatest?**

I can imagine there’s a point in the fact that computing IA queries requires much CPU and RAM - more energy. That has to come from somewhere.

If we can use IA to stabilize things, to offload stress and over-working-hours globally for everyone?

If we can pay for it now, and IA is cheaper, let’s have everyone “who would have gotten fired” continue to work less time for the same money. And some goes to “computing consumption”.

This is a choice.

To be made. Now. By every single one of us.

# Yes, this is my serious face.

- Modern AI is a very powerful thing.
- It is a new, and different kind of technology.  
(The Genie is out of the bottle and here to stay)
- **Who would you like to be in control, and understanding of what?**
- Who knows how stuff works?  
How to...? etc?



What would you consider the right speed, time and ways to socially include this, to **benefit all humans?**

# **(Re)writing history? (Semi-)automatically?**

*“If history keeps repeating itself:*

*Why not come up with a better history,  
which we’d like to see and (re)live?”*

# Local AIs: Pros and Cons

## Pros

+ Consistent. Only changing when you want

+ Ability to train and focus on your data. (and load large local data faster)

+ Possibly lower costs  
(& energy footprint)

+ Stay in control.  
(over privacy, security, tech-understanding)

- **Stable = less updates = use as long as you like**

## Cons

- Need Linux/Sysadmin knowhow

- If it's down, your phone rings

- Need for high-power computing equipment

# Conclusion

*With local AI setups, there is a chance to **gaining understanding** of these new technologies.*

*Allowing us to **actively participate** in what is happening in the (digital) world.*

Long term. Regional. Stable. 😊

# Links

- [Ollama](#)
- [Ollama Documentation](#)
- [Open-WebUI](#)
- [Huggingface Models](#)
- [Whisper \(Speech-to-Text\)](#)
- [Proxmox Virtual Environment \(PVE\)](#)
- [Virtualbox](#)
- [South Park: Deep learning \(S26\)](#)
- [Demoscene Music Video: Mechasm \(by Fairlight, 2023\)](#)
- [The A.I. Dilemma \(March 9, 2023\), by “Center for Humane Technology”](#)

# - The End -

- Author: **Joan De Smith** Long-Term Computing Expert
- eMail: **IA @ ArkThis.com**

which

is

reel?

