IUNO XR | CULTURE | AI

UPDATED JUNE 2026

LOCAL AI

● runs locally

Language models on **your own** machine: private, offline, no subscription.
Your data never leaves the device.

01 Why local?

We all know AI from the cloud: you type something into a web window, the text goes to a remote server, an answer comes back.
Local AI flips that around. The model sits as a file on your computer or home server and runs on your own hardware: no remote server, no login, no one reading along. Today it is set up for free in just a few minutes.

Cloud AI

- Your inputs leave your machine
- Stored, logged, possibly used for training
- Needs internet & usually a subscription
- Tied to one provider

Local AI

+ Everything stays on your machine
+ No logs, no training on your data
+ Runs offline once downloaded, then free
+ Models freely swappable, self-determined

02 Get started in 3 steps

1

Install

Ollama: a simple app or for the terminal. Free for Windows, Mac, Linux.

2

Download

Download an open model once. After that it works fully offline.

3

Use

Ask questions, write text, generate code, all without the cloud.

HOW TO START A MODEL WITH OLLAMA

```
# Install Ollama (app or installer from ollama.com).
# After that, identical on Windows, macOS and Linux:
ollama run qwen3.5:4b
>>> Explain local AI in two sentences.
Local AI means: a language model runs directly on your
device instead of in the cloud. Your data stays on the
machine, private, offline and without a subscription.
```

03 Which model for which machine?

Your hardware	Recommendation	What for
~8 GB RAM, no dedicated GPU (older laptop)	`qwen3.5:2b`, `qwen3.5:4b` ~2.7 to 3.4 GB	Fast answers, summarizing, simple texts. Astonishingly capable for the size.
16 GB RAM or newer Apple Silicon (M-chip) (modern laptop)	`qwen3.5:9b`, `gemma4:e4b` ~6.6 to 9.6 GB	The sweet spot. Solid for everyday use, research, code, multilingual.
NVIDIA GPU with 16 GB+ VRAM or Apple Silicon with 32 GB+ (workstation)	`gemma4:26b` ~17 GB, 70B models from ~48 GB	Close to cloud quality for demanding tasks and long texts.

04 Privacy: your advantage with local AI

+ **Nothing leaves the device**
  No third party, no transfer abroad, no data-processing agreement needed.
+ **No training on your data**
  Inputs and chat history stay local. You decide what gets saved.
+ **Good for sensitive data**
  Suitable for personal, research, client or patient data.

Note: local does not automatically mean GDPR-compliant. Device security (encryption, access control, backups) remains your responsibility. This is not legal advice.

Co-funded by the European Union
Innovatives Potential KULTUR INP
Senate Department for Culture and Social Cohesion BERLIN
htw. Hochschule für Technik und Wirtschaft Berlin University of Applied Sciences

REVERSE

TOOLS & TERMS

open source

What to start with, and what the jargon means, with zero prior knowledge.

05 The tools

Ollama TERMINAL

The engine in the background. Loads and starts models with one command. The de-facto standard almost everything else builds on.

For: anyone okay with the terminal, also on servers

LM Studio APP

Same idea as Ollama, but as a slick desktop app. Search, download, chat, no command line.

For: the easiest start, no terminal

Open WebUI BROWSER

A ChatGPT-like window in the browser that sits in front. Multiple users, history, file upload, great for a studio.

For: a familiar chat interface, shared use

Obsidian NOTES + DB

A note app built from plain text files that you own. With a plugin, a local model chats with your own knowledge.

For: private "ask my archive" assistance (see RAG)

06 Key terms in a minute

LLM (Large Language Model)
The "brain". It learned patterns from huge amounts of text and predicts the next likely word, one at a time.

Token
Smallest chunks of text (word pieces). Models work in tokens, not letters. "Hello" is ~2 tokens.

Parameters `(e.g. 8B)`
Rough size of the model (8B = 8 billion dials). More = often smarter, but needs more memory.

Embedding
Turns text into numbers so the computer can "measure" similarity. The engine behind search & RAG.

Prompt
Your input or instruction. The clearer and more specific, the better the answer.

Context Window
How much text the model holds "in mind" at once. Larger = it forgets less during a conversation.

Quantization `(Q4, Q8)`
Compresses the model into less memory. Makes it smaller & faster with minimal quality loss.

RAG (Retrieval-Augmented Generation)
Before answering, the model looks things up in your own documents, including the source.

07 What works well locally, what (not yet)?

Strengths of local models

+ Private & offline, ideal for sensitive data, contracts, research, notes
+ Writing, summarizing, translating, brainstorming, code
+ Search your own knowledge (RAG) without uploading it
+ No subscription, no limits, full control & customizability

Where cloud is (still) ahead

– The very largest, smartest models need data-center hardware
– Speed depends on your machine, small devices answer slower
– Up-to-date knowledge only with extras (web access or RAG)
– Some setup needed, but then it is entirely yours

08 Try it yourself

Download Ollama
Scan, install, type your first command, ready in 5 minutes.
`ollama.com/download`

More info and links

`ollama.com/library` browse models
`lmstudio.ai` app with a GUI
`openwebui.com` chat frontend
`huggingface.co` home of open models

Co-funded by the European Union
Innovatives Potential KULTUR INP
Senate Department for Culture and Social Cohesion BERLIN
htw. Hochschule für Technik und Wirtschaft Berlin University of Applied Sciences

Christoph Holtmann, christoph.holtmann@htw-berlin.de

This handout was created with the help of AI.