IUNO XR | CULTURE | AI

STAND JUNI 2026

LOKALE KI

● läuft lokal

Sprachmodelle auf dem **eigenen** Rechner: privat, offline, ohne Abo.
Deine Daten verlassen die Maschine nicht.

01 Warum lokal?

KI kennen wir alle: Man tippt etwas in ein Web-Fenster, der Text geht an einen fremden Server, eine Antwort kommt zurück.
Lokale KI verändert das. Das Modell liegt als Datei auf deinem Computer oder deinem Heimserver und rechnet auf deiner eigenen Hardware: kein fremder Server, kein Login, kein Mitlesen. Das ist heute kostenlos in wenigen Minuten eingerichtet.

Cloud-KI

- Deine Eingaben verlassen den Rechner
- Gespeichert, protokolliert, evtl. fürs Training genutzt
- Braucht Internet & meist ein Abo
- An einen Anbieter gebunden

Lokale KI

+ Alles bleibt auf deiner Maschine
+ Keine Logs, kein Training mit deinen Daten
+ Läuft offline, einmal geladen, dann kostenlos
+ Modelle frei austauschbar, selbstbestimmt

02 In 3 Schritten loslegen

1 Installieren
Ollama: simple App oder fürs Terminal. Gratis für Windows, Mac, Linux.

2 Herunterladen
Ein offenes Modell einmalig herunterladen. Danach komplett offline nutzbar.

3 Nutzen
Fragen stellen, Texte schreiben, Code generieren, ganz ohne Cloud.

SO STARTEST DU EIN MODELL MIT OLLAMA

```
# Ollama installieren (App oder Installer von ollama.com).
# Danach auf Windows, macOS und Linux identisch:
ollama run qwen3.5:4b
>>> Erklär mir lokale KI in zwei Sätzen.
Lokale KI heißt: ein Sprachmodell läuft direkt auf deinem
Gerät statt in der Cloud. Deine Daten bleiben dabei auf der
Maschine, privat, offline und ohne Abo.
```

03 Welches Modell für welchen Rechner?

Deine Hardware	Empfehlung	Wofür
~8 GB RAM, keine eigene Grafikkarte (älteres Notebook)	`qwen3.5:2b`, `qwen3.5:4b` ~2,7 bis 3,4 GB	Schnelle Antworten, Zusammenfassen, einfache Texte. Erstaunlich fähig für die Größe.
mind. 16 GB RAM oder neuerer Apple Silicon (M-Chip) (modernes Notebook)	`qwen3.5:9b`, `gemma4:e4b` ~6,6 bis 9,6 GB	Der Sweet Spot. Solide für Alltag, Recherche, Code, mehrsprachig.
NVIDIA-Grafikkarte ab 16 GB VRAM oder Apple Silicon ab 32 GB (Workstation)	`gemma4:26b` ~17 GB, 70B-Modelle ab ~48 GB	Nah an Cloud-Qualität für anspruchsvolle Aufgaben und lange Texte.

04 Datenschutz: dein Vorteil bei lokaler KI

+ **Nichts verlässt das Gerät**
Kein Drittanbieter, kein Transfer in Drittländer, keine Auftragsverarbeitung nötig.

+ **Kein Training mit deinen Daten**
Eingaben und Chatverläufe bleiben lokal. Du entscheidest, was gespeichert wird.

+ **Gut für sensible Daten**
Geeignet für personenbezogene, Forschungs-, Mandanten- oder Patientendaten.

Hinweis: Lokal heißt nicht automatisch DSGVO-konform. Gerätesicherheit (Verschlüsselung, Zugriffsschutz, Backups) bleibt deine Aufgabe. Dies ist keine Rechtsberatung.

Kofinanziert von der Europäischen Union
Innovatives Potential KULTUR INP
Senate Department for Culture and Social Cohesion BERLIN
htw Hochschule für Technik und Wirtschaft Berlin University of Applied Sciences

RÜCKSEITE

WERKZEUGE & BEGRIFFE

open source

Womit du loslegst, und was die Fachwörter bedeuten, ganz ohne Vorwissen.

05 Die Werkzeuge

Ollama TERMINAL

Die Engine im Hintergrund. Lädt und startet Modelle mit einem Befehl. De-facto-Standard, auf den fast alles andere aufsetzt.

Für: alle, die mit dem Terminal okay sind, auch auf Server

LM Studio APP

Dieselbe Idee wie Ollama, aber als schicke Desktop-App. Modelle suchen, laden, chatten, ohne Kommandozeile.

Für: den einfachsten Einstieg ohne Terminal

Open WebUI BROWSER

Ein ChatGPT-ähnliches Fenster im Browser, das vorne dranhängt. Mehrere Nutzer, Verlauf, Datei-Upload, gut fürs Studio.

Für: vertraute Chat-Oberfläche, geteilte Nutzung

Obsidian NOTIZEN + DB

Notiz-App aus reinen Textdateien, die dir gehören. Mit Plugin plaudert ein lokales Modell mit deinem eigenen Wissen.

Für: private „frag mein Archiv"-Assistenz (siehe RAG)

06 Grundbegriffe in einer Minute

LLM (Large Language Model)
Das „Gehirn". Hat aus enorm viel Text Muster gelernt und sagt Wort für Wort das wahrscheinlich nächste voraus.

Token
Kleinste Texthäppchen (Wortteile). Modelle rechnen in Token, nicht in Buchstaben. „Hallo" sind ~2 Token.

Parameter `(z.B. 8B)`
Grobe Größe des Modells (8B = 8 Mrd. Stellschrauben). Mehr = oft klüger, braucht aber mehr Speicher.

Embedding
Verwandelt Text in Zahlen, damit der Computer Ähnlichkeit „messen" kann. Der Motor hinter Suche & RAG.

Prompt
Deine Eingabe oder Anweisung. Je klarer und konkreter, desto besser die Antwort.

Kontextfenster (Context Window)
Wie viel Text das Modell gleichzeitig „im Kopf" behält. Größer = es vergisst weniger im Gespräch.

Quantisierung (Quantization) `Q4, Q8`
Komprimiert das Modell auf weniger Speicher. Macht es kleiner & schneller bei minimalem Qualitätsverlust.

RAG (Retrieval-Augmented Generation)
Das Modell schlägt vor dem Antworten in deinen eigenen Dokumenten nach, inklusive Quelle.

07 Was geht lokal gut, was (noch) nicht?

Stärken lokaler Modelle

+ Privat & offline, ideal für sensible Daten, Verträge, Forschung, Notizen
+ Schreiben, Zusammenfassen, Übersetzen, Brainstorming, Code
+ Eigenes Wissen durchsuchen (RAG) ohne es hochzuladen
+ Kein Abo, keine Limits, volle Kontrolle & Anpassbarkeit

Wo Cloud (noch) vorn liegt

– Die allergrößten, klügsten Modelle brauchen Rechenzentrums-Hardware
– Tempo hängt an deiner Maschine, kleine Geräte antworten langsamer
– Aktuelles Wissen nur mit Zusatz (Web-Anbindung oder RAG)
– Etwas Einrichtung nötig, dafür gehört es danach ganz dir

08 Jetzt selbst ausprobieren

Ollama herunterladen
Scannen, installieren, ersten Befehl tippen, in 5 Minuten startklar.
`ollama.com/download`

Weitere Infos und Links

`ollama.com/library` Modelle durchsuchen
`lmstudio.ai` App mit Oberfläche
`openwebui.com` Chat-Frontend
`huggingface.co` Heimat offener Modelle

Kofinanziert von der Europäischen Union

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

Dieses Handout wurde mit Unterstützung von KI erstellt.