Die Forschungsgreisslerei

institut: wissenschaft.online.beratung

DIGITALKOMPETENZ · EINFÜHRUNG

LOKALE KI

Schritt für Schritt auf Windows und Mac

> Eine verständliche Einführung in lokale Sprachmodelle – von der Installation bis zum datenschutzbewussten Einsatz.

Für wen ist die Broschüre?
Für Einsteiger:innen, die lokale KI ausprobieren oder sensible Inhalte möglichst auf dem eigenen Gerät verarbeiten möchten.

1

Installieren
Jan oder eine vergleichbare Desktop-App installieren.

2

Modell wählen
Ein passendes lokales Modell herunterladen und im Chat auswählen.

3

Lokal prüfen
Websuche ausschalten und mit einem Offline-Test kontrollieren.

ABSCHNITT 1

Schnellstart: in fünf Schritten zur lokalen KI

Wenn Du zuerst ausprobieren möchtest und Details später lesen willst, beginne hier.

1. **Prüfe Deinen Rechner**
   Betriebssystem, Arbeitsspeicher und – unter Windows – den verfügbaren Grafikspeicher (VRAM).

2. **Installiere eine Desktop-App**
   Für lokale Modelle, zum Beispiel Jan, ausschließlich von der offiziellen Website.

3. **Lade ein Modell**
   In der App ein Modell herunterladen, das laut Größen- oder Kompatibilitätsanzeige zu Deinem Gerät passt. Für den Einstieg ist eine Q4- oder Q5-Quantisierung meist sinnvoll.

4. **Wähle das lokale Modell im Chat aus**
   Deaktiviere Websuche, Cloudmodelle und externe Erweiterungen, wenn nichts nach außen übertragen werden soll.

5. **Mache einen Offline-Test**
   Trenne kurz die Internetverbindung und stelle eine einfache Frage. Antwortet das Modell weiter, läuft die Texterzeugung lokal.

Für den ersten Test

Verwende zunächst einen unverfänglichen Text. Erst wenn Du sicher weißt, welche Funktionen lokal und welche extern arbeiten, solltest Du vertrauliche Inhalte verarbeiten.

Kurz gesagt

Du brauchst keine Programmierkenntnisse. Mit einer Desktop-App wie Jan oder LM Studio kannst Du ein Sprachmodell herunterladen und auf Deinem eigenen Rechner verwenden. Die eigentliche Textverarbeitung kann danach offline erfolgen – vorausgesetzt, Du hast ein lokales Modell ausgewählt und Web-, Cloud- oder andere externe Funktionen ausgeschaltet.

ABSCHNITT 2

Was bedeutet „lokale KI“?

Bei lokaler KI läuft das Sprachmodell auf Deinem eigenen Computer. Deine Eingaben müssen für die Antwort nicht an einen externen KI-Anbieter geschickt werden. Das unterscheidet lokale Modelle von typischen Cloud-Diensten, bei denen die Berechnung auf Servern des Anbieters stattfindet.

Wichtig

Eine lokal installierte KI-App kann trotzdem Cloudmodelle, Websuche oder externe Werkzeuge anbieten. „Die App ist lokal installiert“ bedeutet deshalb nicht automatisch „alle Daten bleiben auf dem Gerät“. Entscheidend ist, welches Modell und welche Zusatzfunktionen in der jeweiligen Sitzung aktiv sind.

Was brauchst Du?

- einen Windows-PC oder einen Mac mit Apple Silicon (M1 oder neuer),
- genügend freien Speicherplatz für App und Modelle,
- Internet für Installation und Modelldownload; danach kann ein lokales Modell meist offline genutzt werden,
- eine App, die lokale Modelle laden kann; in dieser Broschüre steht Jan im Mittelpunkt.

Drei Begriffe, die Du kennen solltest

Modell
Die eigentliche KI, z. B. Gemma, Qwen, Llama oder Mistral.

Quantisierung
Eine speichersparende Variante eines Modells. Q4 oder Q5 ist für lokale Nutzung häufig ein sinnvoller Einstieg.

Kontext
Die Textmenge, die das Modell bei einer Antwort gleichzeitig berücksichtigen kann. Mehr Kontext benötigt meist mehr Speicher.

ABSCHNITT 3

Reicht mein Computer?

Kleinere, quantisierte Modelle laufen auf vielen aktuellen Geräten. Mehr Arbeitsspeicher bzw. Grafikspeicher erlaubt größere Modelle, längere Kontexte und meist flüssigeres Arbeiten.

Nutzung	Windows	Mac (Apple Silicon)
Einstieg	16 GB RAM, ca. 8 GB VRAM	16 GB Unified Memory
Gut nutzbar	32 GB RAM, ca. 12 GB VRAM	24 GB Unified Memory
Sehr gut	64 GB RAM, 16–24 GB VRAM	32–48 GB Unified Memory
Große Modelle	64–128 GB RAM, 24 GB+ VRAM	64 GB+ Unified Memory

Orientierungswerte, keine Mindestanforderungen. Modellarchitektur, Quantisierung, Kontextlänge, Betriebssystem und parallel laufende Programme beeinflussen den tatsächlichen Speicherbedarf. Auch mit 16 GB können kleinere Modelle sinnvoll nutzbar sein.

Der wichtigste Unterschied

Unter Windows ist bei GPU-Beschleunigung häufig der Grafikspeicher (VRAM) der Engpass. Beim Mac teilen CPU und GPU den Unified Memory. Reine RAM-Zahlen von Windows und Mac sind daher nur eingeschränkt vergleichbar.

Windows

Einstellungen > System > Info.

Den dedizierten GPU-Speicher findest Du im Task-Manager unter Leistung > GPU.

Mac

Apple-Menü > Über diesen Mac. Dort siehst Du Chip und Arbeitsspeicher bzw. Unified Memory.

Quelle der Pfade: Stand der Systemmenüs zum Zeitpunkt der Broschüre.

ABSCHNITT 4

Jan installieren

Jan ist eine Desktop-App für lokale Sprachmodelle. Du kannst Modelle herunterladen, laden und in einer Chatoberfläche verwenden, ohne mit Terminalbefehlen arbeiten zu müssen.

Windows

1. Öffne die offizielle Jan-Webseite und lade die Windows-Version herunter.
2. Starte die heruntergeladene .exe-Datei und folge der Installation.
3. Öffne Jan.
4. Wenn Du eine NVIDIA-, AMD- oder Intel-Arc-Grafikkarte verwendest, prüfe in Jan unter Settings > Hardware > GPUs, ob die GPU erkannt und aktiviert ist.

Mac

1. Lade die macOS-Version von Jan herunter.
2. Öffne die .dmg-Datei.
3. Ziehe Jan in den Ordner „Programme“ bzw. „Applications“.
4. Starte Jan.

Hinweis zu älteren Macs

Für die lokale Modellausführung mit aktuellen Jan-Versionen ist Apple Silicon vorgesehen. Auf Intel-Macs ist lokale Modellausführung mit der aktuellen Engine nicht der normale Einsatzweg.

ABSCHNITT 5

Das erste lokale Modell herunterladen

Nach der Installation brauchst Du ein Modell. Jan bietet dafür einen integrierten Hub und zeigt bei vielen Modellen an, ob sie voraussichtlich zu Deinem Rechner passen.

1. **Öffne in Jan den Bereich „Hub“**
2. **Suche nach einem eher kleinen oder effizienten Modell**
3. **Achte auf Hinweise wie „Fits“, „May be slow“ oder „Won’t fit“**
4. **Wähle zunächst eine empfohlene bzw. ausgewogene Quantisierung**
   Häufig Q4 oder Q5.
5. **Starte den Download**
   Modelle können mehrere Gigabyte groß sein.
6. **Wähle das heruntergeladene Modell im Chat aus**

Q4, Q5 oder Q8?

Variante	Speicherbedarf	Für den Einstieg
Q4	eher niedrig	meist ein guter Start
Q5	mittel	oft ebenfalls gut geeignet
Q8	hoch / sehr groß	erst bei ausreichend Speicher sinnvoll

Praktische Faustregel

Wähle zuerst ein Modell, das komfortabel in den Speicher passt. Ein etwas kleineres Modell, das stabil läuft, ist im Alltag oft hilfreicher als ein größeres Modell an der Speichergrenze.

ABSCHNITT 6

Prüfen, ob die Sitzung wirklich lokal läuft

„Lokales Modell“ beschreibt die konkrete laufende Sitzung – nicht die App an sich.

1. **Wähle im Modellmenü ausdrücklich das heruntergeladene lokale Modell aus**
2. **Schalte Websuche und externe Werkzeuge aus**
3. **Trenne WLAN bzw. Netzwerk kurz**
4. **Stelle eine einfache Frage**
   Zum Beispiel: „Erkläre mir in fünf Sätzen, was ein Sprachmodell ist.“
5. **Wird die Antwort weiterhin erzeugt, läuft die Texterzeugung des gewählten Modells lokal**

Was kann trotzdem nach außen kommunizieren?

- Websuche und Webseitenabruf benötigen eine Internetverbindung.
- Cloudmodelle von externen Anbietern rechnen nicht auf Deinem Gerät.
- MCP-Server, Plugins oder andere Erweiterungen können zusätzliche Verbindungen herstellen.
- Updates, Telemetrie oder andere App-Funktionen können je nach Software ebenfalls Netzwerkzugriffe verursachen.

> **Datenschutz-Merksatz: Prüfe vor sensiblen Eingaben Modell, Websuche und aktivierte Werkzeuge.**

ABSCHNITT 7

Feineinstellungen: wenige Regler genügen

Für den Anfang solltest Du nur Einstellungen verändern, deren Wirkung Du nachvollziehen kannst. Temperatur, Kontextgröße oder GPU-Layer verändern die Nutzung eines vorhandenen Modells; sie sind kein Fine-Tuning.

Einstellung	Startwert	Wirkung
Temperatur	0,2–0,3 sachlich; ca. 0,7 ausgewogen	Niedriger = fokussierter und berechenbarer; höher = mehr Variation.
Kontextgröße	zunächst ca. 8.192 Tokens	Mehr Kontext erlaubt längere Chats/ Dokumente, benötigt aber mehr RAM/ VRAM.
GPU Layers	Automatik bzw. so hoch wie stabil	Mehr Auslagerung auf die GPU kann beschleunigen, benötigt aber VRAM.
Top P / Top K	Modell-Standard	Für Einsteiger:innen meist kein Grund zum Ändern.
Repeat Penalty	Standard	Erst ändern, wenn das Modell auffällig stark wiederholt.
Threads / Batch / Flash Attention	Automatik / Standard	Technische Leistungsregler; nur bei konkreten Problemen anpassen.

Drei einfache Profile

Einsatz	Temperatur
Fakten, Zusammenfassungen, Fachtexte	etwa 0,1–0,3
Allgemeiner Chat und Beratungsmaterialien	etwa 0,5–0,7
Ideen und kreative Varianten	etwa 0,8–1,0; variabler, aber häufig weniger verlässlich

Wenn ein Modell nicht lädt

Reduziere zuerst die Kontextgröße oder verwende eine kleinere bzw. stärker quantisierte Modellversion. Temperatur ist für Speicherprobleme normalerweise nicht der entscheidende Regler.

ABSCHNITT 8

Websuche: praktisch, aber nicht mehr vollständig lokal

Ein lokales Sprachmodell kennt aktuelle Ereignisse nicht automatisch. Eine zugeschaltete Websuche kann aktuelle Informationen und Quellen liefern. Gleichzeitig verlassen Suchanfragen dann den Rechner.

Sinnvoll bei

- aktuellen Nachrichten, Preisen, Softwareversionen oder neuen Veröffentlichungen,
- der Suche nach öffentlich zugänglichen Quellen,
- Themen, bei denen das lokale Modell erkennbar Wissenslücken hat.

Achte dabei auf

- Suchanfragen werden an einen externen Suchdienst übertragen.
- Namen, Gesundheitsdaten, Fallbeschreibungen oder vertrauliche Dokumentinhalte gehören nicht in Websuchanfragen.
- Webseiten können falsch, veraltet oder interessengeleitet sein. Websuche macht eine Antwort nicht automatisch richtig.

Einfache Regel

- **Rein lokale Arbeit:** Websuche aus.
- **Allgemeine Recherche:** Websuche kann an.
- **Sensible berufliche Inhalte:** vertraulichen Text lokal verarbeiten und nur mit abstrahierten, nicht identifizierenden Suchbegriffen recherchieren.

ABSCHNITT 9

Welche Aufgaben eignen sich lokal?

Lokale Modelle sind besonders interessant, wenn Inhalte auf dem eigenen Gerät bleiben sollen und keine tagesaktuellen Informationen nötig sind.

Gut geeignet

- Texte zusammenfassen oder strukturieren
- Formulierungen überarbeiten
- Ideen und Varianten entwickeln
- lokale Dokumente analysieren
- Fragen zu eigenen Unterlagen beantworten (RAG)
- Entwürfe für Übungen, Arbeitsblätter oder Beratungsmaterialien erstellen
- einfache Programmier- oder HTML-Aufgaben bearbeiten

Grenzen beachten

- Kleine Modelle sind bei komplexem Schlussfolgern oder Spezialthemen oft schwächer als große Cloudmodelle.
- Aktuelles Wissen fehlt ohne Websuche.
- Sehr lange Dokumente und große Kontexte benötigen viel Speicher.
- Auch lokale Modelle können falsche Angaben erzeugen. Wichtige Ergebnisse müssen fachlich geprüft werden.

Empfehlung für den Einstieg

Suche nicht sofort nach dem größten Modell. Nimm zuerst ein Modell, das sicher in den Speicher passt, lerne die Bedienung kennen und teste danach schrittweise andere Modelle.

ABSCHNITT 10

Wenn es langsam wird oder nicht funktioniert

Problem	Was Du versuchen kannst
Modell lädt nicht	Kleinere Quantisierung oder anderes Modell wählen; andere speicherintensive Programme schließen.
Antwort ist sehr langsam	Kleineres Modell wählen; Kontext verkürzen; GPU-Aktivierung prüfen.
Speicher ist voll	Nicht benötigte Modelle löschen und ausreichend SSD-Platz freihalten.
Windows nutzt die GPU nicht	Grafiktreiber aktualisieren und in Jan unter Hardware prüfen, ob die GPU aktiviert ist.
Mac wird langsam	Andere speicherintensive Apps schließen; kleineres Modell oder kleineren Kontext wählen.
Antwortqualität ist schwach	Prompt präzisieren oder ein geeigneteres/größeres Modell testen. Größer ist nicht automatisch besser.

Alternativen zu Jan

LM Studio

Grafische Oberfläche mit Modellsuche, lokalem Chat und vielen Einstellmöglichkeiten. Gut, wenn Du verschiedene Modelle vergleichen möchtest.

Ollama

Technischer orientiert und besonders praktisch, wenn lokale Modelle von anderen Programmen oder über eine lokale API angesprochen werden sollen.

Auch bei Alternativen prüfen

Einige Programme unterstützen zusätzlich Cloudmodelle oder externe Funktionen. Der Programmname allein sagt nicht, ob eine konkrete Sitzung lokal ist.

ABSCHNITT 11

Datenschutz bei sensiblen Inhalten

Lokale KI kann externe Datenübertragung stark reduzieren. Datenschutz hängt aber nicht nur davon ab, wo das Modell rechnet. Auch der Rechner, gespeicherte Chats, Backups und Zugriffsrechte gehören zur Betrachtung.

- Prüfe, dass tatsächlich ein lokales Modell ausgewählt ist.
- Deaktiviere Websuche und externe Erweiterungen, wenn keine Daten nach außen gelangen sollen.
- Schütze den Rechner mit Benutzerkonto, Bildschirmsperre, Festplattenverschlüsselung und aktuellen Sicherheitsupdates.
- Kläre, wo Chatverläufe und hochgeladene Dokumente gespeichert werden und wie sie gelöscht werden können.
- Nutze sichere Backups und berücksichtige, dass auch Backups sensible Daten enthalten können.
- Bei beruflicher Verarbeitung personenbezogener oder besonderer Kategorien personenbezogener Daten gelten zusätzlich die rechtlichen und organisatorischen Anforderungen Deiner Einrichtung bzw. Tätigkeit.

Für professionelle Nutzung

„Lokal“ ersetzt keine Datenschutzorganisation. Zugriffsschutz, Aufbewahrung, Löschung, Berechtigungen, Backups und die fachliche Prüfung von Ergebnissen bleiben erforderlich.

ABSCHNITT 12

Die 10-Minuten-Checkliste

- ☐ Ich weiß, ob mein Rechner Windows oder Apple Silicon verwendet.
- ☐ Ich kenne meinen RAM bzw. Unified Memory und – bei Windows – ungefähr meinen VRAM.
- ☐ Ich habe die App von der offiziellen Website installiert.
- ☐ Ich habe ein Modell ausgewählt, das laut App bzw. Speicherbedarf zu meinem Gerät passt.
- ☐ Ich starte mit einer Q4/Q5- oder empfohlenen Quantisierung.
- ☐ Ich habe im Chat ausdrücklich ein lokales Modell ausgewählt.
- ☐ Ich weiß, ob Websuche und externe Werkzeuge ein- oder ausgeschaltet sind.
- ☐ Ich habe einen Offline-Test gemacht.
- ☐ Ich beginne mit unverfänglichen Testtexten.
- ☐ Ich prüfe wichtige Fakten, Formulierungen und fachliche Aussagen selbst.

Empfehlung für den Einstieg

Suche nicht sofort nach dem größten Modell. Nimm zuerst ein Modell, das sicher in den Speicher passt, lerne die Bedienung kennen und teste danach schrittweise andere Modelle.

ABSCHNITT 13

Stand, Quellen und Hinweise

Stand der Broschüre: 25. Oktober 2026. Software, Systemanforderungen und Funktionsumfang ändern sich. Prüfe vor Installation oder sicherheitsrelevanter Nutzung die aktuelle Dokumentation der jeweiligen Anbieter.

Weiterführende Dokumentation

Jan

- Windows Installation
- Mac Installation
- Modelle verwalten
- Desktop Quickstart
- Model Settings
- Web Search
- Local AI Engine / Performance

LM Studio

- System Requirements
- Get started

Ollama

- Download

Hinweise

Haftungsausschluss. Diese Broschüre ist eine allgemeine Orientierung und ersetzt weder die aktuelle technische Dokumentation der genannten Programme noch eine individuelle IT-, Datenschutz- oder Sicherheitsberatung. Trotz sorgfältiger Prüfung kann keine Gewähr für Vollständigkeit, Richtigkeit oder Aktualität übernommen werden. Die Nutzung der beschriebenen Programme, Einstellungen und Vorgehensweisen erfolgt auf eigene Verantwortung. Für Datenverluste, Fehlkonfigurationen, Sicherheitsprobleme, Funktionsstörungen, Schäden an Hard- oder Software sowie für Folgen inhaltlicher Fehler oder unzutreffender KI-Ausgaben wird – soweit gesetzlich zulässig – keine Haftung übernommen.

Hinweis zur KI-Unterstützung. Bei der Erstellung und redaktionellen Überarbeitung dieser Broschüre wurde generative KI unterstützend eingesetzt. Der Inhalt wurde anschließend vom Herausgeber durchgesehen, fachlich geprüft und in dieser Fassung freigegeben. Die menschliche Prüfung ändert nichts daran, dass technische Angaben veralten und Fehler trotz Kontrolle nicht vollständig ausgeschlossen werden können.

Hinweis zu Speicherwerten. Die genannten Werte sind praxisnahe Orientierungswerte und keine garantierten Mindestanforderungen für jedes Modell.

https://www.forschungsgreisslerei.at/digitalkompetenz/

Gratis erhältlich – freiwillig unterstützen

Diese Broschüre ist gratis. Wenn Du ihre Erstellung und Weiterentwicklung unterstützen möchtest, freuen wir uns über eine freiwillige Spende. Den Betrag bestimmst Du selbst.

Scanne den QR-Code mit Deiner Banking-App, um die Überweisungsdaten zu übernehmen.

Impressum

Herausgeber und für den Inhalt verantwortlich:

Medieninhaber & Herausgeber:
Gerhard Hintenberger & Stefan Kühne
Forschungsgreisslerei – Verein zur Förderung von Wissenschaft und Forschung
Nußdorfer Straße 52/6-7, 1090 Wien

Kontakt:
E-Mail: office@forschungsgreisslerei.at
Web: www.forschungsgreisslerei.at

Zentrales Vereinsregister:
ZVR-Zahl: 1198261375

Haftungsausschluss: Diese Broschüre ist eine allgemeine Orientierung und ersetzt weder die aktuelle Dokumentation der genannten Programme noch eine individuelle IT-, Datenschutz- oder Sicherheitsberatung. Für Richtigkeit, Vollständigkeit und Aktualität wird keine Gewähr übernommen.

Urheberrecht: Die Inhalte dieser Broschüre unterliegen dem Urheberrecht. Vervielfältigung und Verbreitung bedürfen der Zustimmung der Herausgeber, soweit sie nicht gesetzlich erlaubt sind.

Stand: 25. Oktober 2026