



Ollama Manager

Erste Schritte

Ein Einstieg für alle, die noch nie mit einer lokalen künstlichen Intelligenz gearbeitet haben.

Pflegedozi · Stand August 2026
Freie Software unter GNU AGPL v3

Inhalt

1. Willkommen	2
Für wen ist dieses Heft?	2
Ein paar Begriffe vorweg	2
2. Was Sie brauchen	3
Der Rechner	3
Zeit	3
3. Installation	4
Schritt für Schritt	4
Programm starten	4
Ist alles in Ordnung?	5
4. Das erste KI-Modell laden	6
Welches Modell soll ich nehmen?	6
So laden Sie es herunter	6
5. Ihr erster Chat	8
Was, wenn die Antwort seltsam ist?	8
6. Die Oberfläche auf einen Blick	10
Die Knopfleiste im Chat	10
7. Wenn etwas nicht klappt	12
Das Programm startet nicht	12
Server läuft: Nein	12
Es kommt keine Antwort	12
Die Antwort dauert sehr lange	12
Der Rechner wird sehr langsam	12
8. Wie es weitergeht	13
Wo liegt was?	13
Weitergeben ist ausdrücklich erlaubt	13

1. Willkommen

Der **Ollama Manager** ist ein Programm, mit dem Sie eine künstliche Intelligenz auf Ihrem eigenen Rechner benutzen können. Sie stellen Fragen, die KI antwortet - ähnlich wie bei ChatGPT, aber mit einem entscheidenden Unterschied:

Warum das wichtig ist

Alles bleibt auf Ihrem Computer. Ihre Fragen, Ihre Dokumente und die Antworten der KI verlassen den Rechner nicht. Es wird nichts an eine Firma im Internet gesendet. Das Programm ist so gebaut, dass es das technisch gar nicht kann - eine eingebaute Sperre verhindert Verbindungen nach draußen.

Gerade wenn Sie mit Patientendaten, Dienstplänen oder anderen vertraulichen Unterlagen arbeiten, ist das der entscheidende Punkt. Sie müssen niemandem vertrauen außer Ihrem eigenen Rechner.

Für wen ist dieses Heft?

Für alle, die den Ollama Manager zum ersten Mal benutzen. Sie brauchen keinerlei Vorkenntnisse. Wir gehen gemeinsam den kürzesten Weg von der Installation bis zur ersten Antwort der KI.

Dieses Heft zeigt bewusst **nur einen Weg** - nämlich den, der funktioniert. Das Programm kann noch viel mehr. Wenn Sie später mehr wissen wollen, finden Sie alles im **ausführlichen Handbuch**.

Ein paar Begriffe vorweg

Begriff	Was damit gemeint ist
KI-Modell	Das eigentliche 'Gehirn'. Eine große Datei, die Sie einmal herunterladen. Danach funktioniert sie ohne Internet.
Ollama	Das Hintergrundprogramm, das die KI-Modelle ausführt. Es läuft unsichtbar mit. Sie müssen es nicht bedienen.
Ollama Manager	Das Fenster, das Sie sehen und bedienen. Darum geht es hier.
Prompt	Fachwort für 'die Frage oder Anweisung, die Sie eintippen'.

2. Was Sie brauchen

Bevor es losgeht, ein kurzer Blick auf die Voraussetzungen.

Der Rechner

	Mindestens	Angenehm
Arbeitsspeicher	8 GB	16 GB oder mehr
Freier Speicherplatz	10 GB	30 GB
Betriebssystem	Linux Mint oder Ubuntu	
Grafikkarte	nicht nötig	beschleunigt die Antworten

Die KI rechnet auf dem Hauptprozessor. Das funktioniert, dauert aber länger als bei den bekannten Diensten im Internet. Rechnen Sie bei einer längeren Antwort mit einer halben bis zwei Minuten. Das ist normal und kein Fehler.

Hinweis

Je weniger Arbeitsspeicher Ihr Rechner hat, desto kleinere KI-Modelle sollten Sie wählen. Welches Modell zu Ihnen passt, steht in Kapitel 4.

Zeit

Planen Sie für die Einrichtung etwa **30 bis 45 Minuten** ein. Der größte Teil davon ist Warten: Das erste KI-Modell ist mehrere Gigabyte groß und muss heruntergeladen werden.

3. Installation

Sie haben einen Ordner mit dem Programm bekommen - entweder als ZIP-Datei oder auf einem USB-Stick. So bringen Sie ihn zum Laufen.

Schritt für Schritt

1. **Ordner entpacken.** Falls Sie eine ZIP-Datei haben: Rechtsklick darauf, dann *Hier entpacken*. Es entsteht ein Ordner namens **ollama-manager-v0.2**.
2. **Terminal öffnen.** Rechtsklick *in* den entpackten Ordner, dann *Im Terminal öffnen*. Es erscheint ein schwarzes Fenster mit Textzeilen. Erschrecken Sie nicht - Sie tippen dort gleich genau einen Befehl ein.
3. **Installation starten.** Tippen Sie den folgenden Befehl ein und drücken Sie die Eingabetaste:

```
bash install.sh
```

1. **Passwort eingeben.** Wenn nach einem Passwort gefragt wird, geben Sie Ihr normales Anmeldepasswort ein. **Beim Tippen sehen Sie nichts** - keine Punkte, keine Sternchen. Das ist so gewollt. Tippen Sie einfach und drücken Sie Eingabe.
2. **Warten.** Die Installation läuft durch und meldet am Ende *Installation abgeschlossen*.

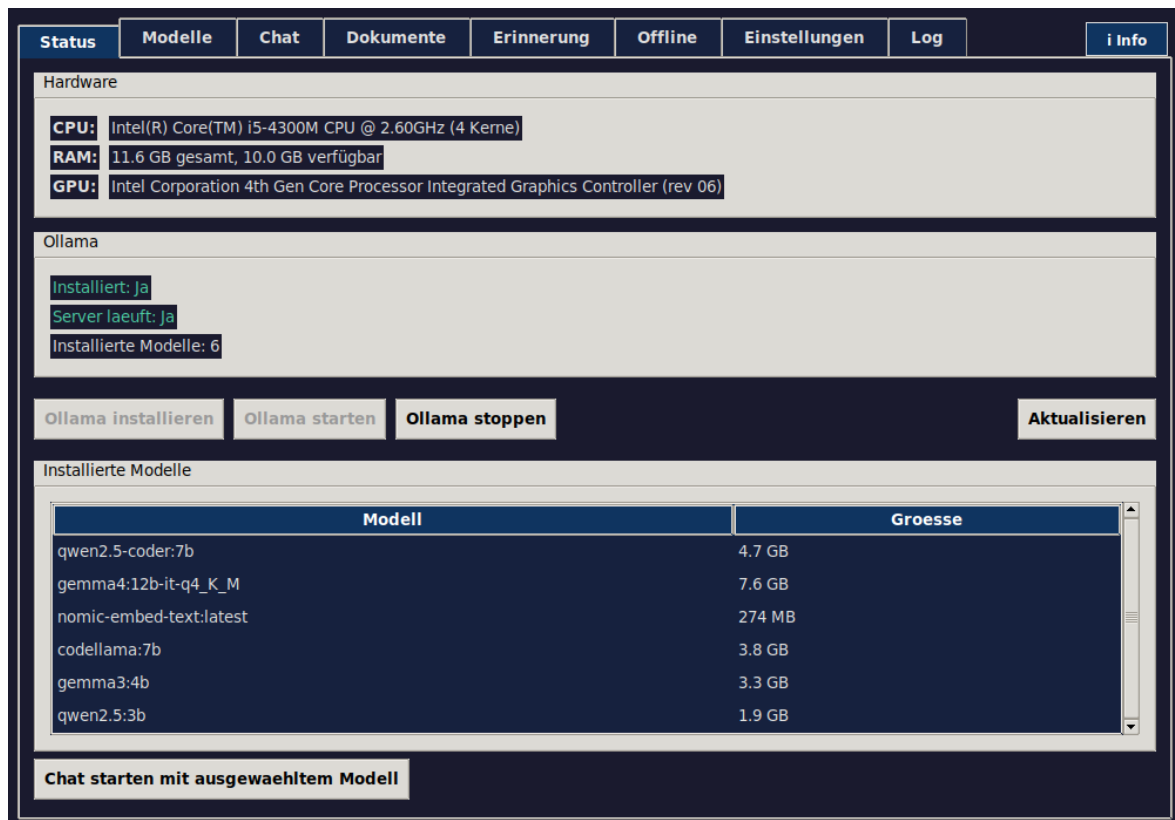
Tipp

Sie müssen im Terminal nichts verstehen. Es ist nur der Weg, das Installationsprogramm zu starten. Danach brauchen Sie es nicht mehr.

Programm starten

Ab jetzt geht es wie bei jedem anderen Programm: Startmenü öffnen, nach **Ollama Manager** suchen, anklicken.

Das Fenster sieht so aus:



Der Tab **Status** - hier sehen Sie, ob alles bereit ist.

Ist alles in Ordnung?

Achten Sie im Bereich **Ollama** auf drei Zeilen:

- **Installiert: Ja** - das Hintergrundprogramm ist da.
- **Server läuft: Ja** - es ist gestartet und bereit.
- **Installierte Modelle** - hier steht am Anfang eine 0. Das ist richtig so; das Gehirn fehlt noch.

Achtung

Steht bei *Server läuft* ein **Nein**, klicken Sie auf den Knopf **Ollama starten** und danach auf **Aktualisieren**.

4. Das erste KI-Modell laden

Das Programm ist installiert, aber es fehlt noch die KI selbst. Die holen Sie sich jetzt - einmalig, danach funktioniert sie ohne Internet.

Welches Modell soll ich nehmen?

Es gibt viele. Für den Anfang empfehlen wir eines der beiden ersten:

Modell	Größe	Wofür	Braucht
qwen2.5:3b	1,9 GB	Guter Allrounder zum Anfangen. Antwortet schnell.	8 GB RAM
gemma3:4b	3,3 GB	Etwas ausführlicher, versteht auch Bilder.	8 GB RAM
qwen2.5-coder:7b	4,7 GB	Für Programmieraufgaben.	16 GB RAM
gemma4:12b	7,6 GB	Sehr gut, aber langsam auf kleinen Rechnern.	16 GB RAM

Empfehlung

Fangen Sie mit **qwen2.5:3b** an. Es ist klein, schnell und läuft auf jedem Rechner. Weitere Modelle können Sie jederzeit dazuholen - sie stören sich nicht gegenseitig.

So laden Sie es herunter

1. Klicken Sie oben auf den Tab **Modelle**.
2. Suchen Sie in der Liste das gewünschte Modell.
3. Klicken Sie auf **Installieren**.
4. Warten Sie. Der Fortschritt erscheint im Tab **Log**. Je nach Internetverbindung dauert das 5 bis 30 Minuten.

StatusModelleChatDokumenteErinnerungOfflineEinstellungenLogi Info

Verfuegbare Modelle fuer deine Hardware

Gruen = empfohlen fuer dein System | Grau = benoetigt mehr Ressourcen

Filter:

Alle

Chat

Code

Vision

Passend

Modell	Beschreibung	Min. RAM	Status
tinylLama	TinyLLama 1.1B - Minimales Chat-Modell	2 GB	Empfohlen
qwen2.5:0.5b	Qwen2.5 0.5B - Ultra-leicht	2 GB	Installiert
gemma3:270m	Gemma 3 270M - Winzig, für Embedded	2 GB	Installiert
gemma3:1b	Gemma 3 1B - Sehr leicht	2 GB	Installiert
phi3:mini	Phi-3 Mini 3.8B - Kompakt & fähig	4 GB	Empfohlen
qwen2.5:3b	Qwen2.5 3B - Gute Balance	4 GB	Installiert
gemma2:2b	Gemma 2 2B - Google's kompaktes Modell	4 GB	Empfohlen
gemma4:e2b	Gemma 4 E2B - Agentic für Handys	3 GB	Installiert
llama3.2:3b	LLama 3.2 3B - Meta's neuestes kompaktes Modell	6 GB	Empfohlen
mistral	Mistral 7B - Bewährter Allrounder	6 GB	Empfohlen
qwen2.5:7b	Qwen2.5 7B - Starkes mittleres Modell	8 GB	Installiert
gemma2:9b	Gemma 2 9B - Starke Leistung	8 GB	Empfohlen
gemma3:4b	Gemma 3 4B - Vision-fähig, multilingual	6 GB	Installiert
gemma4:e4b	Gemma 4 E4B - Edge, multimodal + Audio	6 GB	Installiert
llama3.1:8b	LLama 3.1 8B - Meta's solides Modell	8 GB	Empfohlen
gemma3:12b	Gemma 3 12B - Starke Mittelklasse, Vision	10 GB	Installiert

Ausgewaehltes Modell installieren

Empfohlene installieren

Aus dem Netz suchen

Eigene Variante erstellen

Der Tab **Modelle** - hier holen Sie sich KI-Modelle und löschen sie wieder.

Hinweis

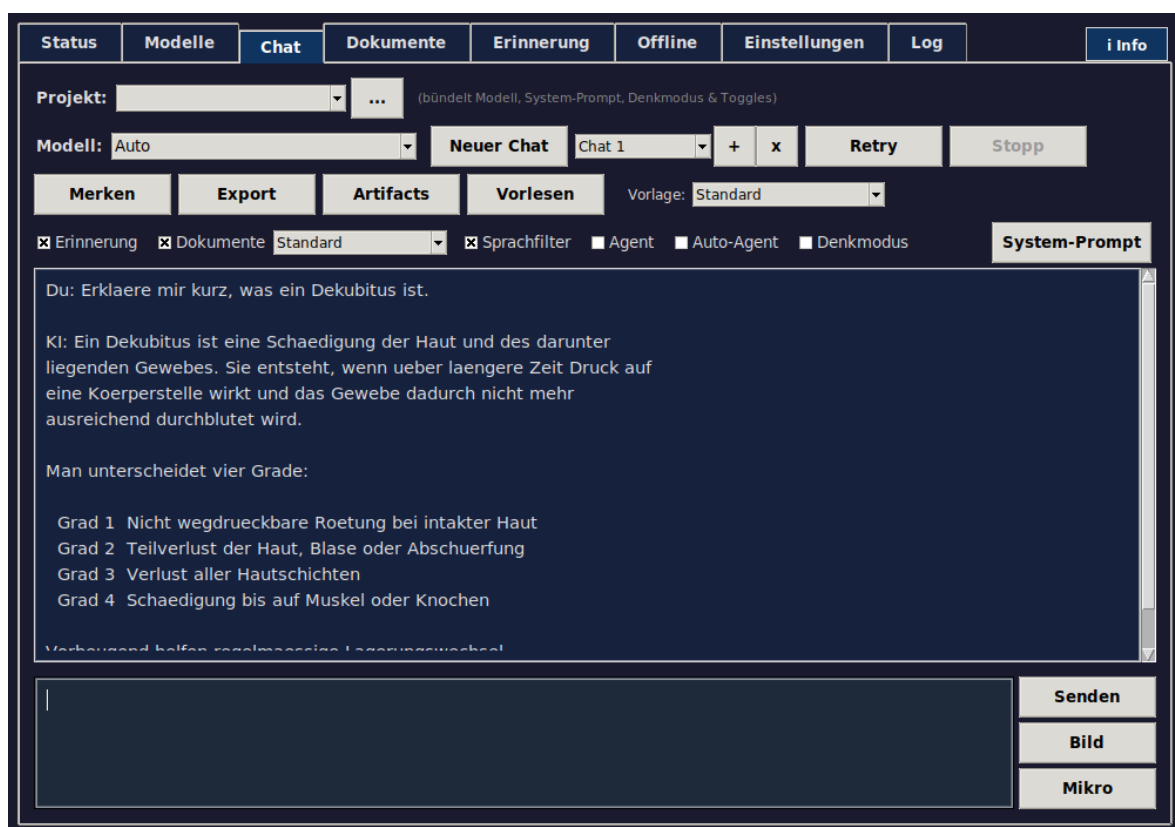
Während des Downloads können Sie den Rechner normal weiterbenutzen. Schließen Sie nur nicht das Programm.

5. Ihr erster Chat

Jetzt kommt der schöne Teil.

1. Klicken Sie oben auf den Tab **Chat**.
2. Wählen Sie bei **Modell** Ihr heruntergeladenes Modell aus.
3. Klicken Sie unten in das große leere Eingabefeld.
4. Tippen Sie eine Frage, zum Beispiel: *Erkläre mir kurz, was ein Dekubitus ist.*
5. Drücken Sie die **Eingabetaste** oder klicken Sie auf **Senden**.

Die Antwort erscheint Wort für Wort. Beim allerersten Mal dauert es etwas länger, weil das Modell in den Arbeitsspeicher geladen wird.



Eine Antwort im Chat.

Zwei nützliche Tasten

Eingabetaste sendet die Frage ab. Wenn Sie eine neue Zeile brauchen, ohne abzuschicken: **Umschalt + Eingabetaste**.

Was, wenn die Antwort seltsam ist?

Kleine KI-Modelle irren sich häufiger als die großen Dienste im Internet. Das ist der Preis dafür, dass alles bei Ihnen bleibt.

- **Fragen Sie genauer.** Je klarer die Frage, desto besser die Antwort.
- **Klicken Sie auf Retry.** Damit wird dieselbe Frage neu beantwortet.
- **Probieren Sie ein größeres Modell,** wenn Ihr Rechner es hergibt.

Bitte unbedingt beachten

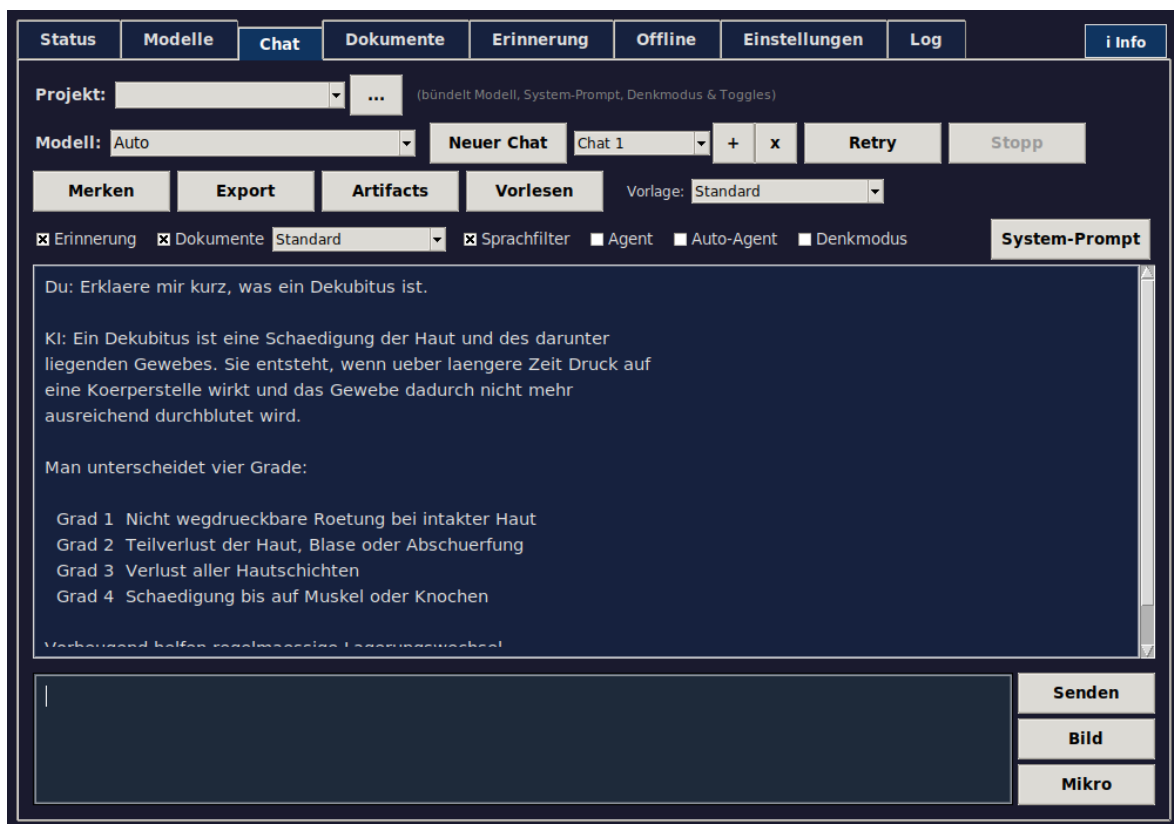
Prüfen Sie fachliche Aussagen immer nach - besonders bei medizinischen oder rechtlichen Themen.
Eine KI kann sehr überzeugend klingen und trotzdem falsch liegen. Sie ist ein Hilfsmittel, keine Quelle.

6. Die Oberfläche auf einen Blick

Oben im Fenster sehen Sie acht Reiter. Sie brauchen am Anfang nur zwei davon - hier trotzdem die Übersicht, damit Sie sich zurechtfinden.

Reiter	Wofür
Status	Ist alles bereit? Welche Modelle sind da?
Modelle	KI-Modelle herunterladen und löschen.
Chat	Hier arbeiten Sie. Der wichtigste Reiter.
Dokumente	Eigene Dateien hinterlegen, damit die KI daraus antworten kann.
Erinnerung	Früher geführte Gespräche nachlesen.
Offline	Alles für den Betrieb ohne Internet.
Einstellungen	Sprache, Stimme, Tastenkombination.
Log	Technische Meldungen. Wichtig bei Problemen.

Die Knopfleiste im Chat



Die Bedienelemente oberhalb des Gesprächsfensters.

Knopf	Was er tut
Neuer Chat	Beginnt ein frisches Gespräch.
Retry	Beantwortet Ihre letzte Frage noch einmal.
Stopp	Bricht eine laufende Antwort ab.
Merken	Speichert eine Information dauerhaft für die KI.
Export	Speichert das Gespräch als Datei.
Vorlesen	Liest die letzte Antwort laut vor.
Mikro	Sie sprechen, statt zu tippen.
Bild	Ein Foto anhängen (nur bei bildfähigen Modellen).

Hinweis

Die Häkchen **Agent**, **Auto-Agent** und **Denkmodus** lassen Sie am Anfang aus. Was sie bewirken, steht im ausführlichen Handbuch.

7. Wenn etwas nicht klappt

Die fünf häufigsten Stolpersteine - mit der Lösung daneben.

Das Programm startet nicht

Starten Sie es einmal aus dem Terminal, dann sehen Sie den Grund:

```
python3 ~/.local/share/ollama-manager/ollama_manager.py
```

Die letzte Zeile der Ausgabe ist der Hinweis. Steht dort etwas mit *ModuleNotFoundError*, wurde die Installation nicht vollständig ausgeführt - dann einfach **bash install.sh** wiederholen.

Server läuft: Nein

Klicken Sie im Tab **Status** auf **Ollama starten**, dann auf **Aktualisieren**. Hilft das nicht, im Terminal:

```
systemctl --user restart ollama
```

Es kommt keine Antwort

- Haben Sie oben ein **Modell** ausgewählt? Ohne Modell passiert nichts.
- Ist im Tab **Status** mindestens ein Modell aufgelistet?
- Schauen Sie in den Tab **Log** - dort steht meist die Ursache.

Die Antwort dauert sehr lange

Das ist normal, besonders beim ersten Mal nach dem Programmstart. Dauert es dauerhaft länger als zwei Minuten, ist das Modell für Ihren Rechner zu groß. Nehmen Sie ein kleineres.

Der Rechner wird sehr langsam

Das Modell belegt zu viel Arbeitsspeicher. Schließen Sie andere Programme oder wechseln Sie auf ein kleineres Modell. Das Programm bremst große Modelle automatisch ab, damit nichts abstürzt - das kann die Antworten aber verkürzen.

Der wichtigste Tipp

Bei jedem Problem lohnt zuerst ein Blick in den Tab **Log**. Dort steht fast immer im Klartext, was schiefgelaufen ist.

8. Wie es weitergeht

Sie können jetzt mit der KI arbeiten. Das Programm kann aber noch einiges mehr:

- **Eigene Dokumente** hinterlegen und die KI daraus antworten lassen - etwa Pflegestandards oder Leitlinien.
- **Diktieren statt tippen** über den Mikrofon-Knopf.
- **Vorlesen lassen**, wenn die Hände gerade nicht frei sind.
- **Projekte** anlegen, damit die KI je nach Aufgabe anders antwortet.
- **Dateien lesen lassen** - die KI kann selbstständig auf Ihre Dateien zugreifen, immer erst nach Ihrer ausdrücklichen Freigabe.

All das steht im **ausführlichen Handbuch**, das Sie zusammen mit diesem Heft bekommen haben. Sie müssen es nicht von vorne bis hinten lesen - es ist ein Nachschlagewerk.

Ein Rat zum Schluss

Nehmen Sie sich eine Woche Zeit mit dem einfachen Chat, bevor Sie die weiteren Funktionen ausprobieren. Wer die Grundlagen sicher beherrscht, tut sich mit dem Rest deutlich leichter.

Wo liegt was?

Was	Wo
Das Programm	~/local/share/ollama-manager/
Ihre Gespräche und Einstellungen	~/local/share/ollama-manager/data/
Die KI-Modelle	~/ollama/models/

Hinweis

Das Zeichen ~ steht für Ihren persönlichen Ordner. Sichern Sie gelegentlich den Ordner **data** - dort stecken alle Ihre Gespräche und Einstellungen. Die Modelle brauchen Sie nicht zu sichern, die sind jederzeit neu ladbar.

Weitergeben ist ausdrücklich erlaubt

Der Ollama Manager ist **freie Software** unter der GNU AGPL v3. Sie dürfen ihn benutzen, so oft und so lange Sie möchten - und Sie dürfen ihn **weitergeben**, an Kolleginnen und Kollegen, an Ihre Einrichtung, an wen Sie wollen. Es kostet nichts, es gibt keine Anmeldung und keinen Freischaltcode.

Nur wenn Sie den Quelltext verändern *und* die veränderte Fassung anderen zugänglich machen, kommt eine Pflicht dazu: Auch diese Fassung muss dann unter der AGPL stehen und ihren Quelltext mitbringen. Für alle, die das Programm einfach benutzen, ändert sich dadurch nichts. Die

Einzelheiten stehen im Kapitel **Lizenz und Weitergabe** des ausführlichen Handbuchs.

Copyright © 2026 Benjamin Tobies