



Ollama Manager

Handbuch

Vollständige Beschreibung aller Funktionen, der Bedienung und der Problembehebung.

Pflegedozi · Stand August 2026
Freie Software unter GNU AGPL v3

Inhalt

1. Über dieses Handbuch	4
Wie dieses Handbuch geschrieben ist	4
Grundbegriffe	4
2. Installation und Einrichtung	6
Systemvoraussetzungen	6
Installation	6
Was wird zusätzlich installiert?	6
Deinstallation	7
3. KI-Modelle	8
Modelle herunterladen	8
Welches Modell wofür?	8
Speicherbedarf und die eingebaute Bremse	9
Modelle löschen	10
4. Der Chat	11
Aufbau von oben nach unten	11
Die Schalter im Einzelnen	11
Tastenkürzel	12
Mehrere Gespräche gleichzeitig	13
Schnellbefehle	13
Die Befehlspalette	13
Artefakte	13
Bilder	14
5. Projekte	15
Die mitgelieferten Projekte	15
Eigenes Projekt anlegen	15
6. Wissensdatenbanken	16
Wie das funktioniert	16
Mehrere Wissensdatenbanken	16
Ordner einlesen	17

Änderungen nachziehen	17
Suche testen	17
7. Erinnerungen und Aufgaben	18
Dauerhafte Erinnerungen	18
Aufgaben	18
Gesprächsverlauf	19
Bewertungen	19
8. Der Agent	20
Zwei Betriebsarten	20
Das Ampelsystem	20
Was der Agent nicht darf	20
Ein Ablauf in der Praxis	21
Abbrechen	21
Welche Modelle können das?	21
9. Fremdwerkzeuge einbinden (MCP)	22
Einen Server einrichten	22
Sicherheit bei Fremdwerkzeugen	22
Warum Sie Werkzeuge auswählen sollten	22
10. Sprache: Vorlesen und Diktieren	24
Vorlesen	24
Diktieren	24
11. Einstellungen	26
12. Schnellzugriff	28
Tastenkombination	28
Symbol in der Taskleiste	28
Automatisch starten	28
13. Datenschutz und Sicherheit	29
Die Netzwerksperre	29
Keine Nutzungsstatistik	29
Selbst nachprüfen	29
Wo liegen Ihre Daten?	29

Sicherung	30
14. Problembehebung	31
Das Programm startet nicht	31
Server läuft: Nein	31
Ollama bricht mitten in der Antwort ab	31
Die Antworten sind schlecht	31
Der Auto-Agent ruft keine Werkzeuge auf	32
Die Web-Suche findet nichts	32
Vorlesen funktioniert nicht	32
Die Tastenkombination reagiert nicht	32
Selbsttest	32
15. Lizenz und Weitergabe	33
Was Sie dürfen	33
Was Sie im Gegenzug tun müssen	33
Kurz zusammengefasst	33
Die KI-Modelle gehören nicht dazu	34
Keine Gewährleistung	34
Wo der vollständige Text steht	34
16. Anhang	35
Alle Tabs auf einen Blick	35
Wichtige Verzeichnisse	35
Nützliche Befehle	35
Wenn gar nichts mehr geht	36

1. Über dieses Handbuch

Dieses Handbuch beschreibt den **Ollama Manager** vollständig. Es ist ein Nachschlagewerk - Sie müssen es nicht von vorne bis hinten lesen. Schlagen Sie das Kapitel auf, das Sie gerade brauchen.

Zuerst das andere Heft

Wenn Sie das Programm zum ersten Mal benutzen, lesen Sie bitte zuerst das Heft **Erste Schritte**. Es führt Sie in einer knappen Stunde von der Installation bis zur ersten Antwort. Dieses Handbuch setzt voraus, dass das Programm läuft.

Wie dieses Handbuch geschrieben ist

- **Fachbegriffe werden erklärt**, wenn sie zum ersten Mal auftauchen.
- **Befehle zum Abtippen** stehen in grauen Kästen.
- **Farbige Kästen** heben Wichtiges hervor - siehe unten.

Hinweis

So sieht eine ergänzende Information aus.

Tipp

So sieht ein Ratschlag aus, der Ihnen Arbeit spart.

Achtung

So sieht eine Warnung aus. Bitte aufmerksam lesen.

Wichtig

So sieht ein Hinweis aus, dessen Missachtung Schaden anrichtet.

Grundbegriffe

Begriff	Bedeutung
KI-Modell	Die eigentliche künstliche Intelligenz - eine Datei von 1 bis 8 GB. Wird einmal heruntergeladen und läuft danach ohne Internet.
Ollama	Das Hintergrundprogramm, das KI-Modelle ausführt. Läuft als Dienst mit und wird nicht direkt bedient.
Prompt	Die Frage oder Anweisung, die Sie eingeben.

Begriff	Bedeutung
System-Prompt	Eine Dauer-Anweisung an die KI, die bei jeder Frage unsichtbar mitgeschickt wird. Zum Beispiel: 'Antworte immer auf Deutsch.'
Kontext	Alles, was die KI bei einer Frage gleichzeitig 'im Kopf' hat: Ihre Frage, den bisherigen Gesprächsverlauf, gefundene Dokumente. Der Platz dafür ist begrenzt.
Token	Die Einheit, in der KI-Modelle rechnen - ungefähr eine Silbe. Wichtig nur, weil der Kontext in Token gemessen wird.
Agent	Betriebsart, in der die KI selbst Dateien lesen oder Befehle ausführen darf - immer erst nach Ihrer Freigabe.
Werkzeug	Eine Fähigkeit, die die KI im Agent-Modus benutzen kann, etwa 'Datei lesen' oder 'im Web suchen'.

2. Installation und Einrichtung

Systemvoraussetzungen

	Minimum	Empfohlen	Anmerkung
Arbeitsspeicher	8 GB	16-32 GB	Bestimmt, wie große Modelle laufen
Speicherplatz	10 GB	30 GB	Pro Modell 1-8 GB
Prozessor	4 Kerne	8 Kerne	Bestimmt das Tempo
Grafikkarte	keine	NVIDIA	Beschleunigt deutlich
System	Linux Mint / Ubuntu		Andere ungetestet

Installation

Im entpackten Programmordner ein Terminal öffnen und eingeben:

```
bash install.sh
```

Das Installationsprogramm erledigt folgendes:

- Es prüft, ob Python und die Fensterbibliothek vorhanden sind, und installiert sie bei Bedarf nach.
- Es kopiert das Programm nach **~/local/share/ollama-manager/**.
- Es prüft anschließend, ob wirklich alle Bestandteile angekommen sind, und bricht ab, falls etwas fehlt.
- Es legt einen Startmenü-Eintrag und den Terminal-Befehl **ollama-manager** an.
- Es installiert die Zusatzpakete aus **requirements.txt**.

Tipp

Das Installationsprogramm ist gefahrlos mehrfach ausführbar. Wenn Sie unsicher sind, ob etwas fehlt: einfach noch einmal laufen lassen. Ihre Daten bleiben dabei unberührt.

Was wird zusätzlich installiert?

Paket	Wofür	Ohne das Paket
chromadb, pypdf	Wissensdatenbank	Der Tab Dokumente bleibt leer
faster-whisper, sounddevice	Spracheingabe	Der Mikro-Knopf meldet einen Hinweis
piper-tts	Sprachausgabe	Vorlesen ist inaktiv
pystray, Pillow	Symbol in der Taskleiste	Kein Tray-Symbol

Paket	Wofür	Ohne das Paket
pynput	Tastenkombination	Kein globaler Hotkey

Das Programm startet auch ohne diese Pakete. Die betroffene Funktion meldet sich dann selbst mit einem Hinweis, statt abzustürzen.

Deinstallation

```
bash ~/.local/share/ollama-manager/uninstall.sh
```

Achtung

Die Deinstallation entfernt das Programm, **nicht** Ihre Daten und nicht die KI-Modelle. Wenn Sie auch die loswerden wollen, müssen Sie **`~/.local/share/ollama-manager/data/`** und **`~/.ollama/`** von Hand löschen.

3. KI-Modelle

Das Modell ist das Gehirn. Ohne Modell antwortet nichts. Dieses Kapitel erklärt, welche es gibt, welches wofür taugt und wie Sie sie verwalten.



Der Tab Modelle.

Modelle herunterladen

- 1. Tab **Modelle** öffnen.
- 2. Gewünschtes Modell in der Liste suchen. Die Filter oben helfen dabei.
- 3. Auf **Installieren** klicken.
- 4. Fortschritt im Tab **Log** verfolgen.

Alternativ im Terminal, wenn Sie ein Modell kennen, das nicht in der Liste steht:

```
ollama pull modellname
```

Welches Modell wofür?

Modell	Größe	Stärke	Bilder	Werkzeuge
qwen2.5:3b	1,9 GB	Allrounder, schnell	nein	ja
gemma3:4b	3,3 GB	Ausführlich	ja	nein

Modell	Größe	Stärke	Bilder	Werkzeuge
qwen2.5-coder:7b	4,7 GB	Programmieren	nein	unzuverlässig
codellama:7b	3,8 GB	Programmieren	nein	nein
gemma4:12b	7,6 GB	Beste Qualität	ja	ja
nomic-embed-text	274 MB	nur Wissensdatenbank	-	-

Nicht jedes Modell kann, was es verspricht

In der Spalte **Werkzeuge** steht, ob ein Modell den Agent-Modus beherrscht. Verlassen Sie sich dabei nicht auf die Angabe von Ollama selbst: **qwen2.5-coder:7b** meldet Werkzeug-Unterstützung, gibt die Aufrufe aber als Text aus, statt sie auszuführen. Das Programm erkennt das und weist Sie darauf hin.

Für den Agent-Modus ist **qwen2.5:3b** die verlässliche Wahl.

Speicherbedarf und die eingebaute Bremse

Ein Modell muss vollständig in den Arbeitsspeicher passen, zusätzlich zum Kontext. Reicht der Speicher nicht, beendet das Betriebssystem den Ollama-Dienst - mitten in der Antwort, ohne Vorwarnung.

Damit das nicht passiert, hat das Programm einen **Speicherschutz**. Er vergleicht die Modellgröße mit dem freien Arbeitsspeicher und begrenzt bei Bedarf automatisch den Kontext. Auf einem Rechner mit reichlich Speicher greift er nicht ein.

Wenn er eingreift, sagt er es Ihnen im Chat:

```
Kontext auf 4096 begrenzt: gemma4:12b braucht 7.0 GB, verfügbar sind 10.0 GB.
```

Einstellbar unter **Einstellungen** → **Speicherschutz**: *automatisch* (empfohlen), *aus* oder ein fester Wert.

Zusatzschutz bei wenig Speicher

Zusätzlich empfehlenswert, wenn Sie wenig Arbeitsspeicher haben: Ollama daran hindern, mehrere Modelle gleichzeitig vorzuhalten. Dazu einmalig im Terminal:

```
sudo mkdir -p /etc/systemd/system/ollama.service.d
sudo tee /etc/systemd/system/ollama.service.d/override.conf <<'ENDE'
[Service]
Environment="OLLAMA_MAX_LOADED_MODELS=1"
ENDE
sudo systemctl daemon-reload && sudo systemctl restart ollama
```

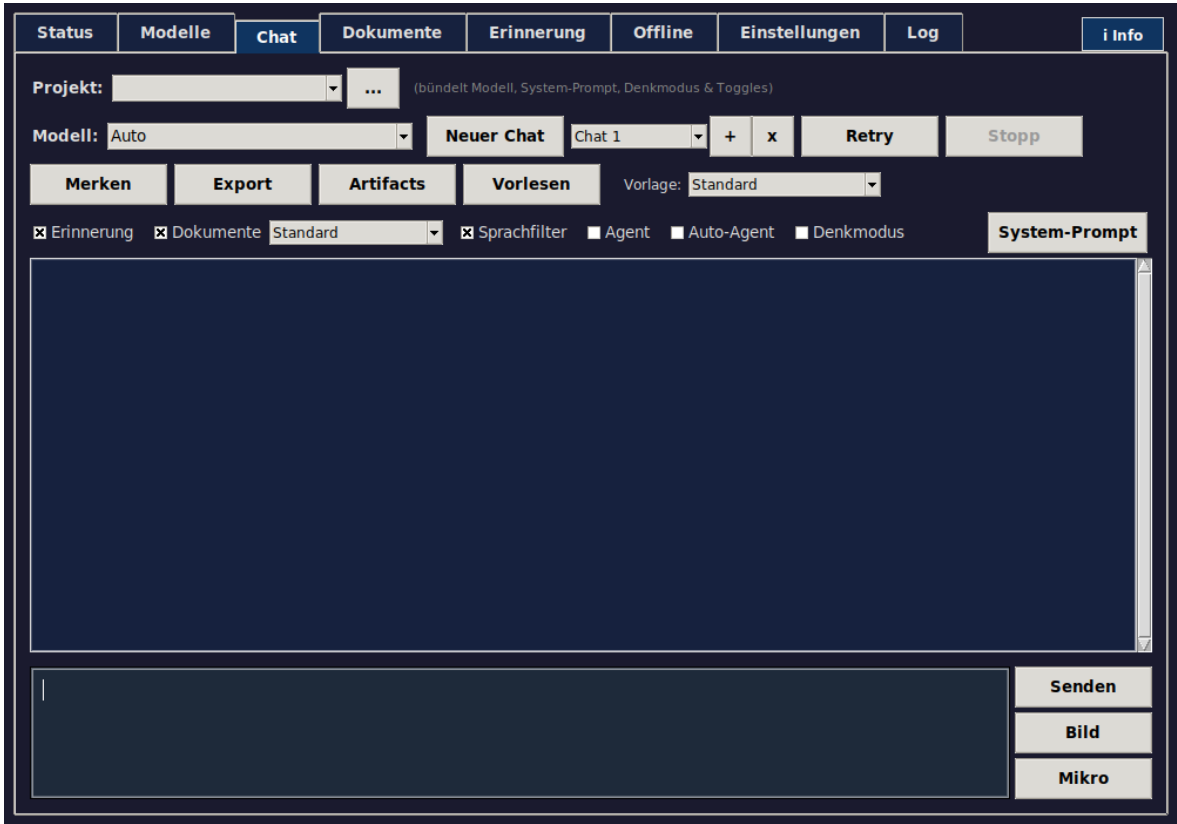
Der Preis: Beim Wechsel des Modells wird das vorherige entladen und muss beim Zurückwechseln neu geladen werden.

Modelle löschen

Im Tab **Modelle** das Modell auswählen und auf **Löschen** klicken. Der Speicherplatz wird sofort frei. Sie können das Modell jederzeit erneut herunterladen.

4. Der Chat

Der Tab **Chat** ist der Arbeitsplatz. Hier läuft alles zusammen.

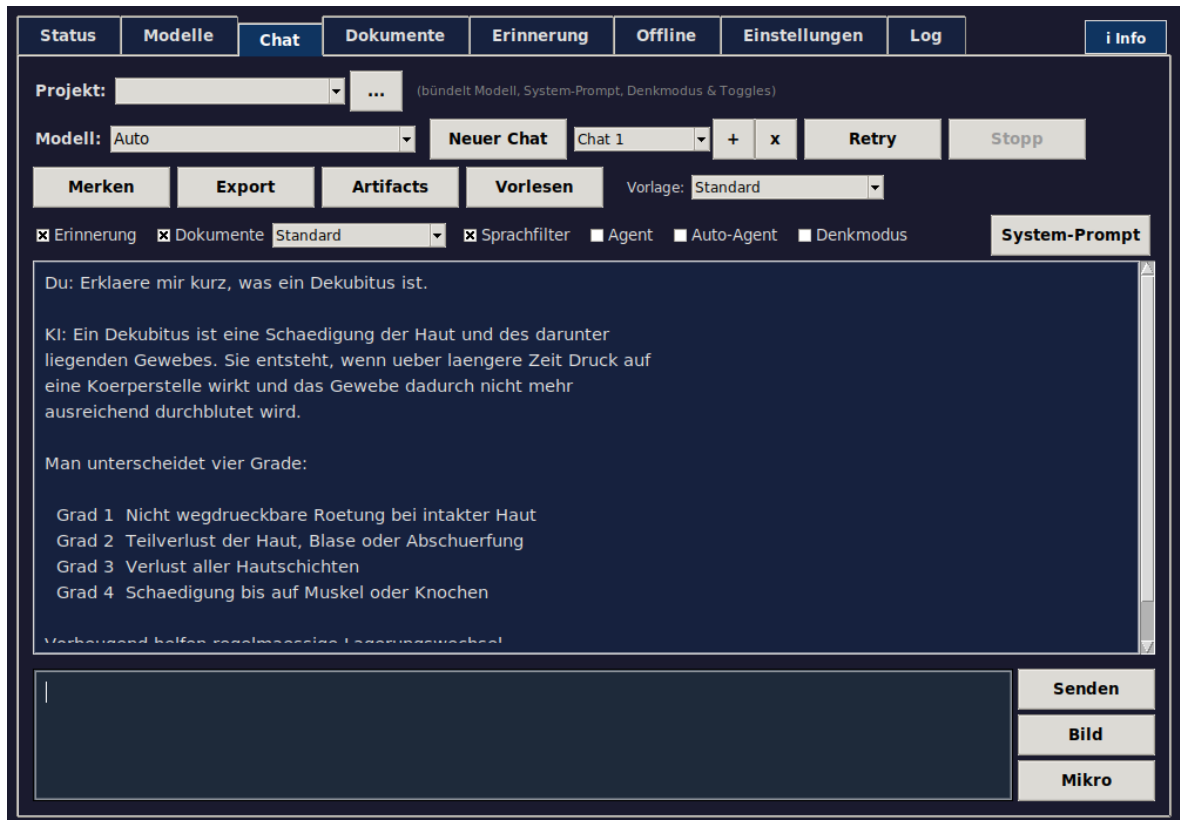


Der Chat vor der ersten Frage.

Aufbau von oben nach unten

Zeile	Inhalt
1. Projekt	Bündelt Modell, System-Prompt und Schalter. Siehe Kapitel 5.
2. Modell	Welche KI antwortet. <i>Auto</i> wählt automatisch.
3. Aktionen	Merken, Export, Artifacts, Vorlesen, Vorlage.
4. Schalter	Erinnerung, Dokumente, Sprachfilter, Agent, Denkmodus.
5. Verlauf	Das Gespräch.
6. Eingabe	Ihre Frage, mit Senden, Bild und Mikro.

Die Schalter im Einzelnen



Die Schalterleiste.

Schalter	Wirkung	Empfehlung
Erinnerung	Sucht in früheren Gesprächen nach Passendem und gibt es der KI mit.	an
Dokumente	Durchsucht Ihre Wissensdatenbank und gibt Fundstellen mit.	an, wenn befüllt
Sprachfilter	Entfernt Floskeln aus der Antwort ('Das ist eine gute Frage...').	an
Agent	Blendet eine Werkzeugleiste ein, die Sie selbst bedienen.	nach Bedarf
Auto-Agent	Die KI wählt ihre Werkzeuge selbst. Siehe Kapitel 8.	bewusst einsetzen
Denkmodus	Das Modell denkt sichtbar vor, bevor es antwortet. Nur bei Modellen, die das können. Kostet Zeit.	aus

Tastenkürzel

Taste	Wirkung
Eingabe	Frage absenden
Umschalt + Eingabe	Neue Zeile, ohne zu senden
Strg + N	Neuer Chat

Taste	Wirkung
Strg + E	Gespräch exportieren
Strg + K	Befehlspalette öffnen

Mehrere Gespräche gleichzeitig

Mit **+** neben dem Chat-Auswahlfeld legen Sie einen zweiten, dritten, beliebig vielen Chat an. Jeder hat seinen eigenen Verlauf. Mit **x** schließen Sie den aktuellen. Die Gespräche überleben einen Programmneustart.

Schnellbefehle

Wenn Sie Ihre Eingabe mit einem Schrägstrich beginnen, wird die Frage automatisch umformuliert:

Befehl	Wirkung
/code	Bittet um Quelltext mit Erklärung
/erkläre	Bittet um eine verständliche Erklärung
/zusammenfassen	Fasst einen längeren Text zusammen
/plan	Erstellt eine nummerierte Schrittfolge
/kritik	Prüft einen Text kritisch
/übersetzen	Übersetzt
/brainstorm	Sammelt Ideen
/merken	Speichert etwas dauerhaft - ohne die KI zu fragen

Hinweis

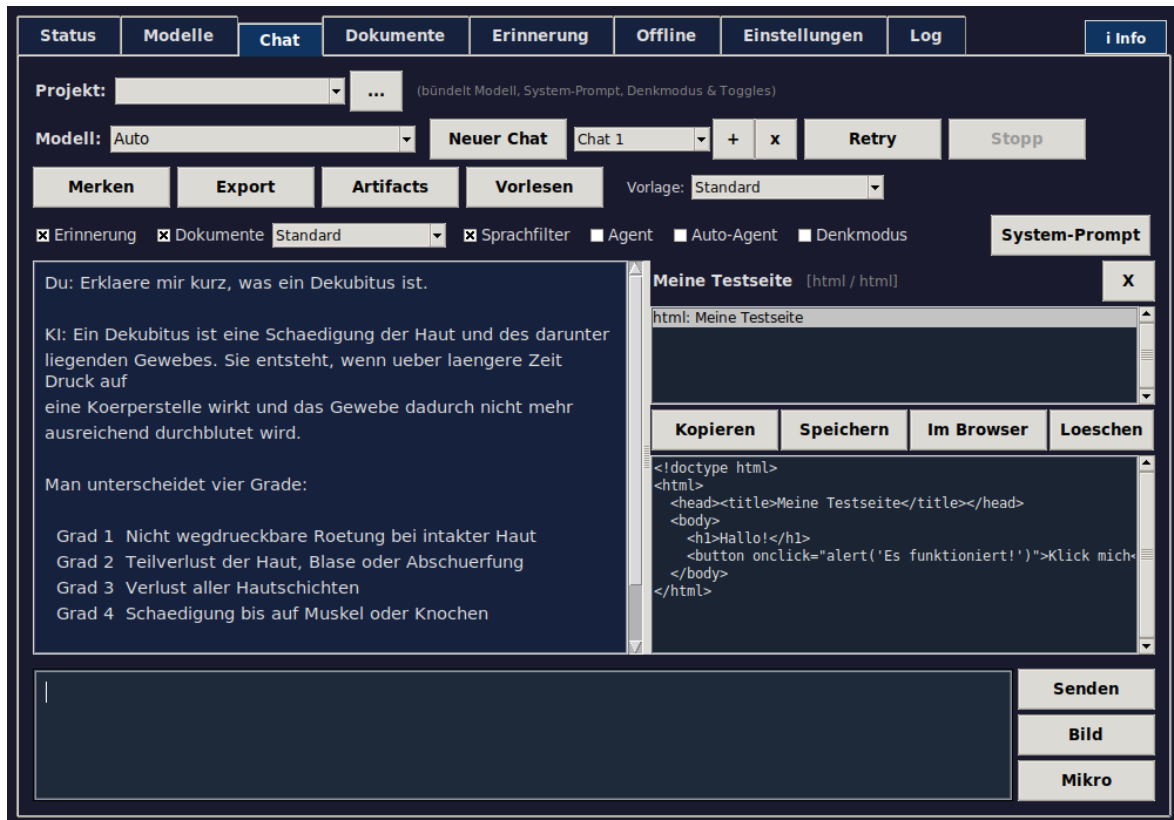
Ein unbekannter Schnellbefehl wird einfach als normale Frage behandelt. Sie können also nichts kaputt machen.

Die Befehlspalette

Strg + K öffnet ein Suchfeld. Tippen Sie ein paar Buchstaben, und es zeigt passende Projekte, Modelle, Schnellbefehle und Tabs. Mit den Pfeiltasten wählen, mit Eingabe ausführen, mit Escape schließen. Der schnellste Weg, wenn Sie wissen, was Sie wollen.

Artefakte

Liefert die KI Quelltext oder eine HTML-Seite, erkennt das Programm das und legt es als **Artefakt** rechts neben dem Gespräch ab.



Links das Gespräch, rechts die erkannten Artefakte.

Dort können Sie den Inhalt kopieren, als Datei speichern oder eine HTML-Seite direkt im Browser ansehen. Der Bereich klappt von selbst auf, sobald es etwas zu zeigen gibt, und lässt sich über den Knopf **Artifacts** wieder wegklappen.

Bilder

Der Knopf **Bild** hängt ein Foto an die nächste Frage. Das funktioniert nur mit bildfähigen Modellen (siehe Kapitel 3). Steht das Modell auf *Auto*, wählt das Programm selbst ein passendes - und sagt Ihnen Bescheid, wenn keines installiert ist, statt das Bild stillschweigend zu ignorieren.

5. Projekte

Ein Projekt bündelt alle Einstellungen für eine bestimmte Art von Arbeit: welches Modell, welcher System-Prompt, welche Schalter. Ein Klick, und die KI verhält sich passend zur Aufgabe.

Die mitgelieferten Projekte

Projekt	Gedacht für
Allgemein	Alltagsfragen, keine Besonderheiten
Code	Programmieraufgaben
Pflegedozent	Fachliche Fragen aus der Pflege, mit passendem System-Prompt und eingeschalteter Dokumentensuche

Eigenes Projekt anlegen

1. Im Chat auf den Knopf ... neben dem Projekt-Auswahlfeld klicken.
2. Auf **Neu** klicken und einen Namen vergeben.
3. Modell wählen (oder *Auto* lassen).
4. System-Prompt eintragen - die Dauer-Anweisung an die KI.
5. Schalter setzen: Erinnerung, Dokumente, Sprachfilter, Auto-Agent, Denkmodus.
6. Speichern.

Beispiel für einen System-Prompt:

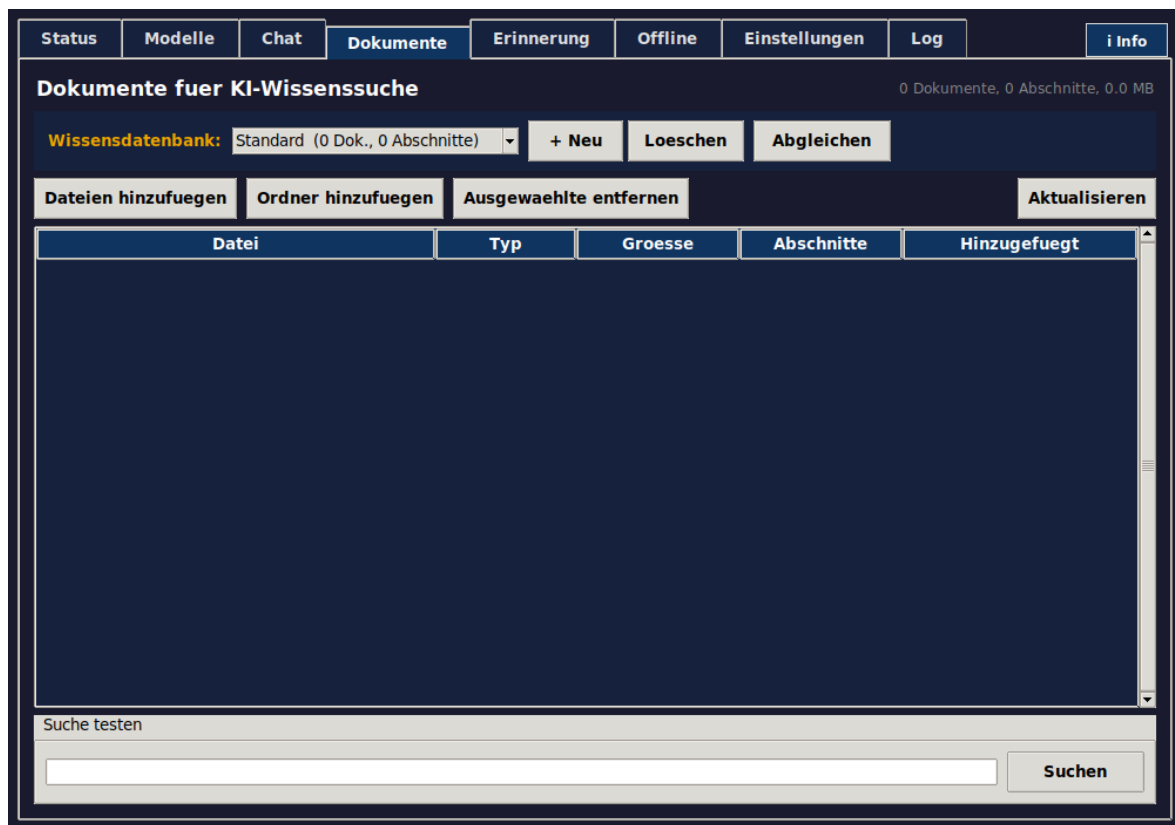
Du bist eine erfahrene Pflegefachkraft. Antworte fachlich korrekt und auf Deutsch. Nenne bei fachlichen Aussagen die Grundlage. Wenn du etwas nicht sicher weißt, sage das ausdrücklich.

Tipp

Der letzte Satz ist der wichtigste. Ohne ihn erfinden kleine Modelle oftmals etwas, das plausibel klingt.

6. Wissensdatenbanken

Sie können eigene Dokumente hinterlegen - Pflegestandards, Leitlinien, Handbücher. Die KI durchsucht sie bei jeder Frage und antwortet aus Ihren Unterlagen statt aus ihrem allgemeinen Wissen.



Der Tab **Dokumente**.

Wie das funktioniert

Beim Hinzufügen wird jedes Dokument in Abschnitte zerlegt. Zu jedem Abschnitt berechnet ein kleines Zusatzmodell einen 'Bedeutungsfingerabdruck'. Stellen Sie später eine Frage, wird auch dafür ein Fingerabdruck gebildet und die ähnlichsten Abschnitte herausgesucht.

Der Vorteil gegenüber einer Wortsuche: Es findet auch dann etwas, wenn Sie andere Wörter benutzen als im Dokument.

Hinweis

Dafür wird das Modell **nomic-embed-text** (274 MB) gebraucht. Das Programm bietet den Download an, sobald Sie das erste Dokument hinzufügen.

Mehrere Wissensdatenbanken

Sie können mehrere getrennte Bestände anlegen - etwa 'Pflegewissen' und 'Projekt Alpha'. Im Chat wählen Sie aus, welcher befragt wird. Das hält die Ergebnisse sauber: Eine Frage zur Pflege

findet dann keine Projektnotizen.

1. Tab **Dokumente** öffnen.
2. Auf **+ Neu** neben der Auswahl klicken und einen Namen vergeben.
3. Dateien oder einen ganzen Ordner hinzufügen.
4. Im Chat neben dem Haken **Dokumente** die gewünschte Datenbank auswählen.

Ordner einlesen

Ordner hinzufügen liest einen ganzen Ordner samt Unterordnern ein. Versteckte Ordner und nicht lesbare Dateitypen werden dabei übersprungen. Vor dem Start sehen Sie, wie viele Dateien es sind.

Unterstützte Dateitypen:

```
.pdf .txt .md .rst .csv .log .py .js .html .xml .json .yaml .yml .ini .cfg .conf
```

Änderungen nachziehen

Der Knopf **Abgleichen** prüft alle Dokumente einer Datenbank: Geänderte Dateien werden neu eingelesen, gelöschte aus dem Bestand entfernt. Erkannt wird das an einer Prüfsumme, nicht am Datum - eine nur geöffnete Datei wird also nicht unnötig neu verarbeitet.

Suche testen

Unten im Tab können Sie eine Suche ausprobieren, ohne die KI zu bemühen. Die Treffer erscheinen im Tab **Log** mit Prozentangabe zur Relevanz. Nützlich, um zu prüfen, ob die Datenbank überhaupt das Richtige findet.

Achtung

Große Bestände brauchen Zeit. Rechnen Sie mit etwa einer Sekunde pro Textabschnitt. Ein Ordner mit hundert PDF-Dateien kann eine halbe Stunde beschäftigt sein. Sie können währenddessen weiterarbeiten.

7. Erinnerungen und Aufgaben

The screenshot shows the 'Erinnerung' tab in the Ollama Manager interface. At the top, there is a navigation bar with tabs: Status, Modelle, Chat, Dokumente, **Erinnerung**, Offline, Einstellungen, Log, and i Info. Below the navigation bar, the interface is divided into three main sections:

- Permanente Erinnerungen**: This section displays a table with columns 'Erinnerung', 'Kategorie', and 'Gilt fuer'. Below the table are buttons for '+ Neu', 'Bearbeiten', and 'Loeschen'. The status bar indicates '0 allgemein, 0 modellspezifisch | Feedback: 18+ / 0-'.
- Aufgaben**: This section displays a table with columns 'Aufgabe', 'Prio', and 'Faellig'. Below the table are buttons for '+ Neu', 'Erledigt', and 'Loeschen'. The status bar indicates '0 offen, 0 erledigt'.
- Chat-Erinnerungen**: This section includes a search bar, buttons for 'Suchen', 'Alle anzeigen', 'Aktualisieren', 'Chat fortsetzen', and 'Ausgewählte löschen'. Below these is a table with columns 'Thema', 'Modell', 'Nachrichten', and 'Datum'. The status bar indicates '23 Gespräche, 74 Nachrichten'.

Thema	Modell	Nachrichten	Datum
Wie geht es dir? Wie geht es dir?	qwen2.5:3b	6	27.07.2026 22:53
Wie geht es dir? Wie geht es dir?	qwen2.5:3b	4	27.07.2026 22:50

Der Tab Erinnerung.

Dauerhafte Erinnerungen

Das sind Tatsachen über Sie, die die KI bei **jeder** Frage mitbekommt - unabhängig vom Gespräch. Zum Beispiel:

- Ich arbeite als Altenpfleger.
- Antworte mir immer knapp, höchstens fünf Sätze.
- Ich unterrichte angehende Pflegefachkräfte.

Anlegen über den Knopf **Merken** im Chat oder über **+ Neu** im Tab Erinnerung. Eine Erinnerung kann für alle Modelle gelten oder nur für ein bestimmtes.

Tipp

Sparsam sein. Jede Erinnerung belegt Platz im Kontext, der dann für die eigentliche Frage fehlt. Fünf gute Erinnerungen sind besser als dreißig mittelmäßige.

Aufgaben

Eine einfache Aufgabenliste mit Priorität und Fälligkeitsdatum. Offene Aufgaben werden der KI als Kontext mitgegeben - Sie können also fragen: *Was steht heute an?*

Gesprächsverlauf

Jedes Gespräch wird automatisch gespeichert und ist durchsuchbar. Bei eingeschaltetem Haken **Erinnerung** sucht das Programm bei jeder Frage nach passenden früheren Gesprächen und gibt sie mit.

Bewertungen

Nach jeder Antwort erscheinen zwei kleine Knöpfe zum Bewerten. Die Bewertungen werden gespeichert und im Tab Erinnerung gezählt. Sie helfen Ihnen einzuschätzen, welches Modell für Ihre Art von Fragen taugt.

8. Der Agent

Bitte aufmerksam lesen

Dieses Kapitel beschreibt die mächtigste und zugleich heikelste Funktion. Bitte vollständig lesen, bevor Sie sie einsetzen.

Im Agent-Modus kann die KI nicht nur reden, sondern **handeln**: Dateien lesen, Ordner durchsuchen, Befehle ausführen, Webseiten abrufen, Dateien schreiben.

Zwei Betriebsarten

	Agent	Auto-Agent
Wer entscheidet?	Sie	die KI
Bedienung	Knopfleiste im Chat	Haken setzen, fertig
Geeignet für	gezielte einzelne Aktionen	mehrstufige Aufgaben
Voraussetzung	jedes Modell	werkzeugfähiges Modell

Das Ampelsystem

Jedes Werkzeug hat eine Risikostufe. Danach richtet sich, ob Sie gefragt werden:

Stufe	Bedeutung	Werkzeuge
Grün	Wird ohne Rückfrage ausgeführt. Reines Lesen.	Datei lesen, Ordner auflisten
Gelb	Wird angezeigt und protokolliert.	Suchen, Befehl ausführen, Code ansehen, Webseite abrufen
Rot	Fragt jedes Mal nach Ihrer Zustimmung.	Datei schreiben, Datei ändern

Was der Agent nicht darf

Fest eingebaut, nicht abschaltbar:

- Zugriff nur in Ihrem persönlichen Ordner und in **/tmp**. Systemordner sind gesperrt.
- Gesperrt sind außerdem alle Pfade, die nach Zugangsdaten aussehen: **.ssh**, **.gnupg**, **.env**, **credentials**, **.password**, **.key** und ähnliche.
- Befehle nur aus einer festen Liste harmloser Kommandos (ls, cat, df, free, uname, date, du, find, grep und einige mehr).
- Ausdrücklich gesperrt: sudo, rm, chmod, kill, curl, wget, ssh und alles, was Programme startet.
- Keine Sonderzeichen in Befehlen, mit denen sich mehrere Kommandos verketteten ließen.

- Zehn Sekunden Zeitlimit pro Befehl.

Ein Ablauf in der Praxis

1. Sie setzen den Haken **Auto-Agent** und wählen ein werkzeugfähiges Modell.
2. Sie fragen: *Welche Dateien liegen in meinem Dokumente-Ordner?*
3. Die KI ruft das Werkzeug **Ordner auflisten** auf - grün, also sofort.
4. Sie sehen im Chat: **[Werkzeug: list_directory Risiko: green]** und darunter das Ergebnis.
5. Die KI formuliert daraus eine Antwort.

Verlangt die Aufgabe ein rotes Werkzeug, erscheint ein Fenster mit Werkzeugname und allen Parametern. Erst nach Ihrem **Ja** passiert etwas.

Abbrechen

Der Knopf **Stopp** beendet einen laufenden Agent-Vorgang sofort - vor der nächsten Runde, vor dem nächsten Werkzeug und mitten in einer laufenden Antwort.

Welche Modelle können das?

Der Auto-Agent braucht ein Modell mit echter Werkzeug-Unterstützung. Verlässlich ist **qwen2.5:3b**; **gemma4:12b** kann es ebenfalls, braucht aber viel Speicher. Ein ungeeignetes Modell führt nicht zum Absturz - es beantwortet die Frage einfach normal und sagt Ihnen, dass es keine Werkzeuge kann.

9. Fremdwerkzeuge einbinden (MCP)

MCP steht für *Model Context Protocol*. Es ist eine Vereinbarung, mit der fremde Programme der KI zusätzliche Werkzeuge anbieten können - etwa Zugriff auf eine Versionsverwaltung oder eine Datenbank.

Hinweis

Das ist eine Funktion für Fortgeschrittene. Für den normalen Betrieb brauchen Sie sie nicht.

Einen Server einrichten

1. **Einstellungen** → **MCP-Server** öffnen.
2. Auf **+ Neu** klicken.
3. Name frei wählen, Befehl und Argumente eintragen.
4. Auf **Speichern und verbinden** klicken.

Beispiel für einen Dateisystem-Server:

Feld	Wert
Name	dateien
Befehl	npx
Argumente	-y @modelcontextprotocol/server-file-system /home/benutzer/Dokumente

Sicherheit bei Fremdwerkzeugen

Hier läuft fremder Programmcode. Deshalb strenger als bei den eigenen Werkzeugen:

- **Kein MCP-Werkzeug ist jemals grün.** Auch reines Lesen wird angezeigt.
- Standardstufe ist **gelb**.
- **Rot**, sobald der Server ein Werkzeug als verändernd meldet.
- Ein Werkzeug ohne laufenden Server gilt als **rot**.

Warum Sie Werkzeuge auswählen sollten

Jedes angebotene Werkzeug wird bei **jeder** Frage mitgeschickt - auch wenn es gar nicht gebraucht wird. Das kostet Platz im Kontext:

	Zeichen pro Frage
Die eingebauten neun Werkzeuge	2.934

	Zeichen pro Frage
Ein Dateisystem-Server, alle 14 Werkzeuge	9.055
Derselbe Server, zwei Werkzeuge freigegeben	1.251

Zum Vergleich: Ein kleines Modell wie qwen2.5:3b hat insgesamt etwa 3.000 Zeichen Kontext zur Verfügung. Ein einziger Server ohne Auswahl erstickt so ein Modell.

Tipp

Nutzen Sie den Knopf **Werkzeuge** und geben Sie nur frei, was Sie wirklich brauchen. Die Einstellungen zeigen Ihnen die Summe an und warnen, wenn es zu viel wird.

10. Sprache: Vorlesen und Diktieren

Vorlesen

Der Knopf **Vorlesen** im Chat liest die letzte Antwort laut vor. Nützlich, wenn die Hände nicht frei sind oder das Lesen am Bildschirm anstrengt.

Dafür braucht es zwei Dinge:

- das Programm **piper** (wird mitinstalliert),
- eine **Stimmdatei** in **~/.local/share/ollama-manager/voices/**.

Eine deutsche Stimme holen Sie sich so:

```
mkdir -p ~/.local/share/ollama-manager/voices
cd ~/.local/share/ollama-manager/voices
B=https://huggingface.co/rhasspy/piper-voices/resolve/main/de/de_DE/thorsten/medium
curl -fLO $B/de_DE-thorsten-medium.onnx
curl -fLO $B/de_DE-thorsten-medium.onnx.json
```

Achtung

Eine englische Stimme kann deutsche Laute nicht bilden - Umlaute gehen verloren, das Ergebnis klingt falsch. Achten Sie darauf, dass der Dateiname mit **de_** beginnt.

Stimme und Sprechtempo stellen Sie unter **Einstellungen** → **Sprachausgabe** ein. Der Knopf **Stimme testen** spricht einen Beispielsatz.

Diktieren

Der Knopf **Mikro** nimmt auf und schreibt das Gesprochene in das Eingabefeld. Die Erkennung läuft lokal - nichts wird verschickt.

1. Auf **Mikro** klicken.
2. Ein Fenster zeigt einen Countdown. Sprechen Sie.
3. Nach Ablauf wechselt die Anzeige auf *Verarbeite Aufnahme*.
4. Der erkannte Text erscheint im Eingabefeld - Sie können ihn vor dem Absenden noch ändern.

Hinweis

Beim ersten Mal nach dem Programmstart dauert es etwa zehn Sekunden länger: Das Erkennungsmodell wird in den Speicher geladen. Das Fenster sagt Ihnen das.

Einstellbar unter **Einstellungen** → **Spracheingabe**:

Einstellung	Bedeutung
Erkennungsmodell	tiny schnell und ungenau · base guter Kompromiss · small/medium genauer, langsamer
Aufnahmedauer	5 bis 60 Sekunden
Sprache	Voreingestellt Deutsch

11. Einstellungen

StatusModelleChatDokumenteErinnerungOfflineEinstellungenLogi Info

Einstellungen

Sprachausgabe (Vorlesen)

Stimme

de_DE-thorsten-medium.onnx [Deutsch]

Datei...

Sprechtempo

1,00

hoeher = langsamer

Stimme testen

Spracheingabe (Mikro)

Erkennungsmodell

base

groesser = genauer, aber langsamer

Aufnahmedauer

12

Sekunden

Sprache

de

Spracherkennung bereit.

Speicherschutz

Kontextgrenze

automatisch

schuetzt vor Abstuerzen bei grossen Modellen

Arbeitsspeicher: 11.6 GB gesamt, 9.9 GB frei.
Begrenzt wird derzeit: gemma4:12b-it-q4_K_M -> 4096

Schnellzugriff (Hotkey und Tray)

Offnet das Quick-Prompt-Fenster, ohne dass der Manager laufen muss. Beides gehoert deshalb in den Autostart der Sitzung.

Sprachausgabe und Spracheingabe.

StatusModelleChatDokumenteErinnerungOfflineEinstellungenLogi Info

Schnellzugriff (Hotkey und Tray)

Offnet das Quick-Prompt-Fenster, ohne dass der Manager laufen muss. Beides gehoert deshalb in den Autostart der Sitzung.

Tastenkombination

<ctrl>+<alt>+<space>

Uebernehmen

Testen

z.B. <ctrl>+<alt>+<space>

Beim Anmelden starten

☒ Hotkey

☒ Tray-Icon

Jetzt

Hotkey starten

Hotkey stoppen

Tray starten

Tray stoppen

Hotkey laeuft nicht.
Tray laeuft nicht.
Sitzungstyp x11 - globale Hotkeys funktionieren hier.

MCP-Server (Fremdwerkzeuge)

Bindet fremde Werkzeug-Server ein (Dateisystem, Git, Datenbanken). Werkzeuge fremder Server sind nie automatisch erlaubt - mindestens gelb, veraendernde rot.

Server	Status	Werkzeuge	Befehl
--------	--------	-----------	--------

+ Neu

Bearbeiten

Werkzeuge

Loeschen

Verbinden

Werkzeug-Schemas: 2934 Zeichen eigene + 0 Zeichen aus MCP = 2934 Zeichen pro Anfrage.

Schnellzugriff und MCP-Server.

Alle Änderungen werden sofort gespeichert. Es gibt bewusst keinen Speichern-Knopf, den man vergessen kann.

Die Einstellungen liegen in einer Datei:

```
~/.local/share/ollama-manager/data/settings.json
```

12. Schnellzugriff

Beide Funktionen dieses Kapitels öffnen das **Quick-Prompt** - ein kleines Fenster für eine schnelle Frage, ohne den ganzen Manager zu starten.

Tastenkombination

Voreingestellt ist **Strg + Alt + Leertaste**. Änderbar unter **Einstellungen** → **Schnellzugriff**.

Schreibweise: Sondertasten in spitze Klammern, verbunden mit Pluszeichen.

```
Richtig: <ctrl>+<alt>+<space>  
Richtig: <ctrl>+<alt>+k  
Falsch: <ctrl>+<alt>+space
```

Tipp

Der Knopf **Testen** wartet zwanzig Sekunden auf Ihren Tastendruck und sagt Ihnen, ob er angekommen ist. Nutzen Sie ihn - manche Kombinationen fängt die Arbeitsumgebung selbst ab, bevor sie beim Programm ankommen.

Symbol in der Taskleiste

Das Tray-Symbol sitzt **unten rechts in der Taskleiste**, neben Uhr und Lautstärke - **nicht** im Programmfenster. Ein Rechtsklick öffnet ein Menü mit Quick-Prompt, Manager öffnen und Beenden.

Häufige Verwechslung

Ist das Programmfenster maximiert, verdeckt es die Taskleiste - das Symbol scheint dann zu fehlen. Fenster verkleinern, dann ist es wieder da.

Automatisch starten

Die beiden Haken bei **Beim Anmelden starten** sorgen dafür, dass Tastenkombination und Symbol nach jeder Anmeldung bereitstehen - auch ohne geöffneten Manager. Genau dafür sind sie gedacht.

13. Datenschutz und Sicherheit

Das Programm ist von Grund auf so gebaut, dass Ihre Daten den Rechner nicht verlassen. Dieses Kapitel erklärt, wie das durchgesetzt wird.

Die Netzwerksperre

Beim Start wird eine Sperre aktiviert, die alle Verbindungen ins Internet blockiert. Erlaubt sind nur Verbindungen zum eigenen Rechner - dort läuft die KI.

Ausnahmen gibt es nur an drei Stellen, und alle drei stoßen **Sie** bewusst an:

- Herunterladen eines KI-Modells,
- das Werkzeug **Web-Suche** im Agent-Modus,
- das Werkzeug **Webseite abrufen** im Agent-Modus.

Keine Nutzungsstatistik

Eine der verwendeten Zusatzbibliotheken sendet normalerweise Nutzungsdaten an ihren Hersteller. Das Programm ersetzt diese Funktion beim Start durch eine wirkungslose Attrappe. Es wird nichts gesendet.

Selbst nachprüfen

Sie müssen das nicht glauben. Während Sie chatten, im Terminal:

```
ss -tnp | grep -E 'ollama|python'
```

Es dürfen nur Verbindungen zu **127.0.0.1** erscheinen - das ist Ihr eigener Rechner. Externe Adressen dürfen nicht auftauchen.

Wo liegen Ihre Daten?

Was	Wo	Inhalt
Gespräche	data/memory.db	alle geführten Chats
Erinnerungen	data/persistent_memory.db	Erinnerungen, Aufgaben, Bewertungen
Projekte	data/projects.db	Ihre Projekteinstellungen
Wissensdatenbank	data/chromadb/	Ihre Dokumente
Einstellungen	data/settings.json	Konfiguration
KI-Modelle	~/.ollama/models/	die Modelle selbst

Alles unterhalb von **~/.local/share/ollama-manager/**, sofern nicht anders angegeben.

Sicherung

Beenden Sie das Programm und sichern Sie den Datenordner:

```
cp -r ~/.local/share/ollama-manager/data ~/sicherung-$(date +%F)
```

Tipp

Die KI-Modelle brauchen Sie nicht zu sichern - die sind jederzeit neu ladbar. Der Datenordner dagegen enthält alles, was unersetzlich ist.

14. Problembehebung

Der erste Griff

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

Das Programm startet nicht

Aus dem Terminal starten, dann wird der Grund sichtbar:

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

Meldung	Ursache und Lösung
ModuleNotFoundError	Installation unvollständig. bash install.sh wiederholen.
no display name	Kein Grafikbildschirm verfügbar. Das Programm braucht eine Bildschirmsitzung.
database is locked	Das Programm läuft bereits ein zweites Mal. Alle Fenster schließen.

Server läuft: Nein

```
systemctl --user restart ollama  
# oder, falls systemweit installiert:  
sudo systemctl restart ollama
```

Ollama bricht mitten in der Antwort ab

Der Arbeitsspeicher ist ausgegangen; das Betriebssystem hat den Dienst beendet. Nachprüfen lässt sich das so:

```
journalctl -u ollama -n 30 | grep -i 'oom\|killed'
```

Abhilfe, in dieser Reihenfolge:

- **Einstellungen** → **Speicherschutz** auf *automatisch* stellen.
- Ein kleineres Modell verwenden.
- Ollama auf ein geladenes Modell begrenzen (siehe Kapitel 3).
- Andere Programme schließen.

Die Antworten sind schlecht

Beobachtung	Was hilft
Antwortet auf Englisch	System-Prompt setzen: 'Antworte immer auf Deutsch.'
Viele Floskeln	Haken Sprachfilter setzen
Erfindet Tatsachen	Größeres Modell; im System-Prompt zum Zugeben von Unwissen auffordern
Ignoriert die Dokumente	Haken Dokumente gesetzt? Richtige Wissensdatenbank gewählt? Suche im Tab Dokumente testen
Vergisst den Gesprächsverlauf	Kontext ist voll. Neuen Chat beginnen oder Erinnerungen ausdünnen

Der Auto-Agent ruft keine Werkzeuge auf

Fast immer liegt es am Modell. Nehmen Sie **qwen2.5:3b**. Erscheint im Chat der Hinweis, das Modell habe den Aufruf *als Text ausgegeben*, ist das Modell für diesen Zweck ungeeignet - unabhängig davon, was Ollama über seine Fähigkeiten meldet.

Die Web-Suche findet nichts

Die Suche läuft über DuckDuckGo. Ohne Internetverbindung oder bei aktivem VPN kann sie fehlschlagen. Die Fehlermeldung im Chat nennt den Grund.

Vorlesen funktioniert nicht

```
which piper
ls ~/.local/share/ollama-manager/voices/
```

Fehlt das Programm oder die Stimmdatei, holen Sie beides nach - siehe Kapitel 10. Unter **Einstellungen** → **Sprachausgabe** steht im Klartext, was fehlt.

Die Tastenkombination reagiert nicht

- Mit dem Knopf **Testen** prüfen.
- Eine andere Kombination probieren - die Arbeitsumgebung könnte sie abfangen.
- Unter Wayland sind globale Tastenkombinationen systemseitig eingeschränkt. Der Sitzungstyp steht in den Einstellungen.

Selbsttest

Das Programm bringt eine Testsammlung mit, die alle Bestandteile prüft:

```
cd ~/.local/share/ollama-manager
python3 -m unittest tests
```

Am Ende muss **OK** stehen. Andernfalls nennt die Ausgabe, welcher Teil nicht funktioniert.

15. Lizenz und Weitergabe

Der Ollama Manager ist **freie Software**. Er steht unter der **GNU Affero General Public License, Version 3** - kurz **AGPL**. Das ist eine der bekannten Lizenzen für offene Software, und sie hat für Sie vor allem angenehme Folgen.

Was Sie dürfen

- Das Programm **benutzen** - privat wie beruflich, ohne Gebühr, ohne Anmeldung, auf beliebig vielen Rechnern.
- Das Programm **weitergeben** - an Kolleginnen und Kollegen, an andere Einrichtungen, an wen Sie möchten.
- In den **Quelltext sehen** und ihn **verändern**. Der komplette Quelltext liegt im Programmordner; es gibt keinen versteckten Teil.

Was Sie im Gegenzug tun müssen

Nur eines, und es betrifft Sie nur dann, wenn Sie das Programm **verändern und die veränderte Fassung anderen zugänglich machen**:

- Die veränderte Fassung muss ebenfalls unter der AGPL stehen.
- Sie muss ihren Quelltext mitbringen.

Das gilt bei der AGPL auch dann, wenn Sie die veränderte Fassung nicht als Datei weitergeben, sondern **über ein Netzwerk anbieten** - etwa als Webdienst. Wer daraus einen Onlinedienst macht, muss seinen Nutzern den Quelltext zugänglich machen.

Kurz zusammengefasst

Darf ich...	
...das Programm im Betrieb einsetzen?	Ja, ohne Einschränkung
...es an Kollegen weitergeben?	Ja
...Geld dafür verlangen?	Ja, auch das erlaubt die AGPL
...es verändern?	Ja
...eine veränderte Fassung geheim halten?	Nur solange Sie sie für sich behalten
...daraus einen Cloud-Dienst machen?	Ja, aber dann mit offenem Quelltext

Warum die AGPL

Warum ausgerechnet diese Lizenz? Weil dieses Programm dafür gebaut wurde, dass Ihre Daten bei Ihnen bleiben. Eine freizügigere Lizenz würde erlauben, daraus einen geschlossenen Cloud-Dienst zu machen - also genau das, wogegen es antritt. Die AGPL schließt diese Lücke.

Die KI-Modelle gehören nicht dazu

Die Lizenz gilt für **dieses Programm**. Die KI-Modelle, die Sie damit herunterladen, sind eigenständige Werke mit eigenen Bedingungen. Die meisten sind frei nutzbar, einige haben Auflagen für den gewerblichen Einsatz. Wenn Sie ein Modell beruflich einsetzen, werfen Sie einmal einen Blick in dessen Lizenz - auf der Seite, von der es stammt.

Keine Gewährleistung

Wie bei freier Software üblich, wird das Programm ohne Gewährleistung bereitgestellt. Es ist sorgfältig getestet und im täglichen Einsatz, aber niemand haftet dafür, dass es für Ihren Zweck taugt. Halten Sie es wie mit jeder Software: Sichern Sie wichtige Daten regelmäßig.

Wo der vollständige Text steht

Der amtliche Lizenztext ist auf Englisch und liegt als Datei **LICENSE** im Programmordner:

```
~/ .local/share/ollama-manager/LICENSE
```

Im Netz finden Sie ihn unter **www.gnu.org/licenses/agpl-3.0**. Dort gibt es auch inoffizielle Übersetzungen und eine gut lesbare Erklärung in einfacher Sprache.

Copyright © 2026 Benjamin Tobies

16. Anhang

Alle Tabs auf einen Blick

Tab	Inhalt
Status	Hardware, Ollama-Zustand, installierte Modelle
Modelle	Modelle laden, löschen, eigene Varianten bauen
Chat	Der Arbeitsplatz
Dokumente	Wissensdatenbanken
Erinnerung	Erinnerungen, Aufgaben, Verlauf, Bewertungen
Offline	Betrieb ohne Internet
Einstellungen	Sprache, Speicherschutz, Schnellzugriff, MCP
Log	Technische Meldungen

Wichtige Verzeichnisse

Pfad	Inhalt
~/.local/share/ollama-manager/	Programm
~/.local/share/ollama-manager/data/	Ihre Daten
~/.local/share/ollama-manager/voices/	Stimmen
~/.ollama/models/	KI-Modelle
~/config/autostart/	Autostart-Einträge

Nützliche Befehle

Befehl	Wirkung
ollama-manager	Programm starten
ollama list	Installierte Modelle anzeigen
ollama pull NAME	Modell laden
ollama rm NAME	Modell löschen
ollama ps	Welche Modelle sind gerade im Speicher?
systemctl --user restart ollama	Dienst neu starten
python3 -m unittest tests	Selbsttest

Wenn gar nichts mehr geht

1. Tab **Log** lesen.
2. Aus dem Terminal starten und die Ausgabe lesen.
3. Zwischengespeicherte Dateien löschen:

```
find ~/.local/share/ollama-manager -name __pycache__ -exec rm -rf {} +
```

1. Installation wiederholen: **bash install.sh**
2. Selbsttest laufen lassen.

Hinweis

Ihre Daten liegen getrennt vom Programm. Eine Neuinstallation löscht weder Gespräche noch Einstellungen.