

DeLoMa – Deine Lokale Maschine

Handbuch

Vollständige Beschreibung aller Funktionen, der Bedienung und der Problembehebung.

Version 0.3 · Stand September 2026

Pflegedozi · Freie Software unter GNU AGPL v3

pflegedozi.org

Inhalt

1. Über dieses Handbuch 4

Der Name 4

Wie dieses Handbuch geschrieben ist 4

Grundbegriffe 4

2. Installation und Einrichtung 6

Systemvoraussetzungen 6

Installation 6

Update vom Ollama Manager 6

Was wird zusätzlich installiert? 7

Deinstallation 7

3. KI-Modelle 9

Modelle herunterladen 9

Welches Modell wofür? 9

Speicherbedarf und die eingebaute Bremse 10

Modelle löschen 11

4. Der Chat 12

Aufbau von oben nach unten 12

Die Schalter im Einzelnen 12

Tastenkürzel 13

Mehrere Gespräche gleichzeitig 13

Schnellbefehle 13

Die Befehlspalette 14

Artefakte 14

Bilder 15

5. Projekte 16

Die mitgelieferten Projekte 16

Eigenes Projekt anlegen 16

6. Wissensdatenbanken 17

Wie das funktioniert 17

Mehrere Wissensdatenbanken 17

Ordner einlesen 18

Änderungen nachziehen . . . 18
Suche testen . . . 18

7. Erinnerungen und Aufgaben . . . 19
Dauerhafte Erinnerungen . . . 19
Aufgaben . . . 19
Gesprächsverlauf . . . 19
Bewertungen . . . 20

8. Der Agent . . . 21
Zwei Betriebsarten . . . 21
Das Ampelsystem . . . 21
Was der Agent nicht darf . . . 21
Ein Ablauf in der Praxis . . . 22
Abbrechen . . . 22
Welche Modelle können das? . . . 22

9. Fremdwerkzeuge einbinden (MCP) . . . 23
Einen Server einrichten . . . 23
Sicherheit bei Fremdwerkzeugen . . . 23
Warum Sie Werkzeuge auswählen sollten . . . 23

10. Sprache: Vorlesen und Diktieren . . . 25
Vorlesen . . . 25
Diktieren . . . 25

11. Einstellungen . . . 26
Grafikkarte . . . 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 29

14. Offline-Betrieb 31

Übertragen 31

Sichern 31

15. Problembehebung 33

Das Programm startet nicht 33

Server läuft: Nein 33

Ollama bricht mitten in der Antwort ab 33

Jedes Modell scheitert mit „Internal Server Error“ 33

Die Antworten sind schlecht 34

Alles ist im Akkubetrieb langsam 34

Der Auto-Agent ruft keine Werkzeuge auf 34

Die Web-Suche findet nichts 34

Vorlesen funktioniert nicht 35

Der Dokumente-Tab bleibt leer 35

Die Tastenkombination reagiert nicht 35

Selbsttest 35

16. Lizenz und Weitergabe 36

Was Sie dürfen 36

Was Sie im Gegenzug tun müssen 36

Kurz zusammengefasst 36

Der Name gehört nicht dazu 36

Die KI-Modelle gehören nicht dazu 37

Keine Gewährleistung 37

Wo der vollständige Text steht 37

17. Anhang 38

Alle Tabs auf einen Blick 38

Wichtige Verzeichnisse 38

Nützliche Befehle 39

Wenn gar nichts mehr geht 39

1. Über dieses Handbuch

Dieses Handbuch beschreibt DeLoMa 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.

Der Name

DeLoMa steht für **Deine Lokale Maschine**. Bis Version 0.2 hieß das Programm „Ollama Manager“. „Ollama“ ist eine Marke von Ollama Inc., und ein unabhängiges Programm sollte nicht so heißen, als wäre es deren Werkzeug. Am Programm selbst hat sich durch den neuen Namen nichts geändert; was sich geändert hat, steht in Kapitel 2 unter Update vom Ollama Manager.

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 20 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.

Begriff	Bedeutung
Prompt	Die Frage oder Anweisung, die Sie eingeben.
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 1–20 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/deloma/`.
- Es legt eine eigene Python-Umgebung (venv) im Programmordner an und installiert dort die Zusatzpakete aus `requirements.txt`. Fehler dabei werden gemeldet, nicht verschluckt.
- Es prüft anschließend, ob wirklich alle Bestandteile angekommen sind, und bricht ab, falls etwas fehlt.
- Es legt einen Startmenü-Eintrag und den Terminal-Befehl `deloma` an.
- Es bietet an, das Einbettungsmodell `nomic-embed-text` für die Wissensdatenbank zu laden, sobald Ollama läuft.

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

Update vom Ollama Manager

Wer den Ollama Manager (Version 0.2) schon installiert hat, muss nicht neu anfangen. Am einfachsten geht es mit der Update-Datei von pflegedozi.org: herunterladen, im Download-Ordner ein Terminal öffnen und eingeben:

```
bash deloma-update-0.3.sh
```

Die Datei enthält das ganze Programm samt Handbüchern und prüft sich vor dem Entpacken selbst. Dasselbe erreicht man aus dem entpackten Vollpaket heraus mit:

```
bash update.sh
```

Das Update macht der Reihe nach:

1. Es sichert Ihre Daten als Archiv nach `~/deloma-sicherung/`.
2. Es verschiebt Gespräche, Erinnerungen, Wissensdatenbanken, Einstellungen und Stimmdateien von `~/.local/share/ollama-manager/` nach `~/.local/share/deloma/`.
3. Es entfernt die alten Startmenü- und Autostart-Einträge und legt sie unter dem neuen Namen neu an.
4. Es installiert die neue Fassung wie oben beschrieben und legt die Handbücher unter `~/.local/share/deloma/handbuch/` ab.

Danach heißt der Terminal-Befehl `deloma` statt `ollama-manager`. Die KI-Modelle liegen bei Ollama und sind vom Update nicht betroffen.

> **Programm vorher schließen**
> Das Update bricht ab, solange das Hauptfenster, das Tray-Symbol oder der Hotkey-Dienst noch laufen. Schließen Sie alle drei und starten Sie das Update dann erneut.

> **Auch install.sh kann das**
> Das Installationsprogramm erkennt eine vorhandene Installation des Ollama Managers und übernimmt die Daten auf dieselbe Weise. Nur die zusätzliche Sicherung als Archiv legt allein `update.sh` an.

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
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/deloma/uninstall.sh
```

Daten bleiben
Die Deinstallation entfernt das Programm, nicht Ihre Daten und nicht die KI-Modelle. Wenn Sie auch die loswerden wollen, müssen Sie `~/.local/share/deloma/` 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	Braucht	Stärke	Bilder	Werkzeuge
qwen2.5:3b	1,9 GB	8 GB	Allrounder, schnell	nein	ja
gemma3:4b	3,3 GB	8 GB	Ausführlich	ja	nein
gemma4:e4b	9,6 GB	16 GB	Sehr gut, versteht Bilder	ja	ja

Modell	Größe	Braucht	Stärke	Bilder	Werkzeuge
qwen2.5-coder:7b	4,7 GB	16 GB	Programmieren	nein	nein
codellama:7b	3,8 GB	16 GB	Programmieren	nein	nein
qwen3:30b-a3b	18 GB	32 GB	Beste Qualität, trotz Größe schnell	nein	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. Geprüft wurde das mit einer echten Werkzeugfunktion; der Messbericht auf pflegedozi.org zeigt, wie.

Groß heißt nicht langsam

`qwen3:30b-a3b` ist knapp viermal so groß wie `qwen2.5-coder:7b` und trotzdem auf demselben Rechner 77 % schneller. Es rechnet pro Wort nur mit 3 seiner 30 Milliarden Parameter. Wer 32 GB Arbeitsspeicher hat, bekommt damit das beste Modell dieser Liste – nur passt dann kein zweites großes Modell gleichzeitig in den Speicher.

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:e4b braucht 9.6 GB, verfügbar sind 12.0 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 die Befehle aus dem folgenden Kasten ausführen. Der Preis: Beim Wechsel des Modells wird das vorherige entladen und muss beim Zurückwechseln neu geladen werden.

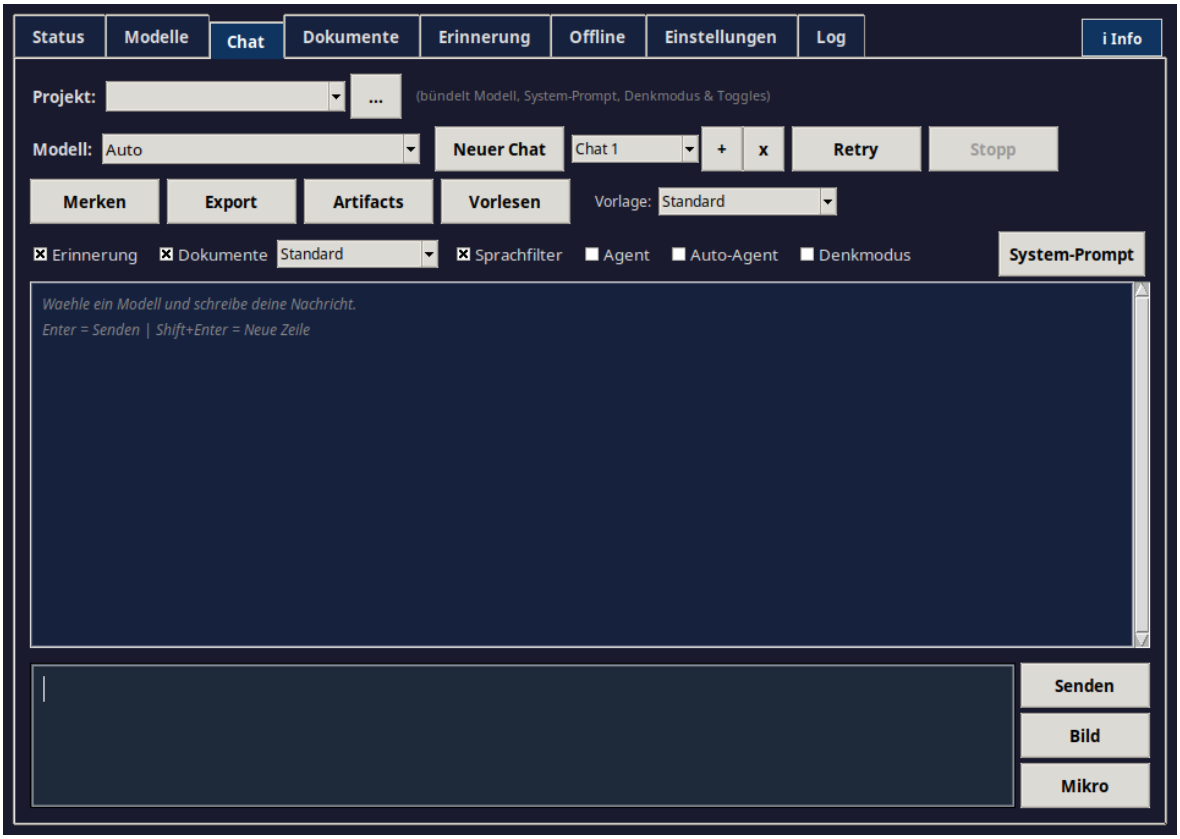
```
sudo mkdir -p /etc/systemd/system/ollama.service.d
sudo tee /etc/systemd/system/ollama.service.d/override.conf <<'ENDE'
[Service]
Environment="OLLAMA_MAX_LOADED_MODELS=1"
ENDE
sudo systemctl daemon-reload && sudo systemctl restart ollama
```

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.

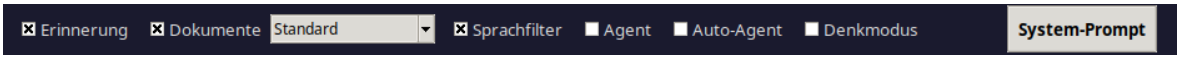


Der Chat vor der ersten Frage.

Aufbau von oben nach unten

Zeile	Inhalt
1. Projekt	Bündelt Modell, System-Prompt und Schalter. Siehe Kapitel 5.
2. Modell	Welche KI antwortet. **Auto** 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

Schalter	Wirkung	Empfehlung
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.	nach Bedarf

> **Der Denkmodus kostet nichts**
> Gemessen mit `qwen3:30b-a3b`: Mit Denken antwortet das Modell knapper, die Geschwindigkeit bleibt gleich. Bei Fachfragen lohnt er sich; bei Smalltalk ist er überflüssig.

Tastenkürzel

Taste	Wirkung
Eingabe	Frage absenden
Umschalt + Eingabe	Neue Zeile, ohne zu senden
Strg + N	Neuer Chat
Strg + E	Gespräch exportieren
Strg + K	Befehlspalette öffnen
Strg + 1 bis Strg + 7	Zwischen den Tabs wechseln

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:

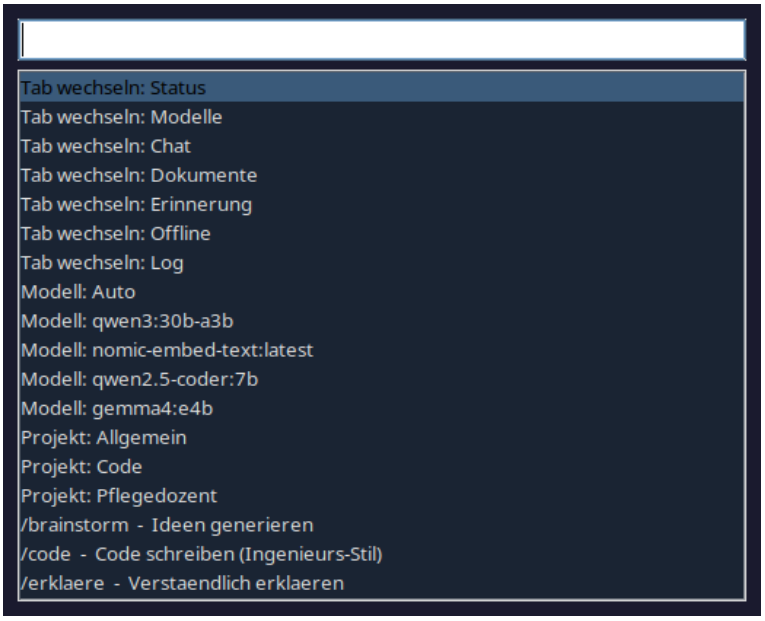
Befehl	Wirkung
/code	Bittet um Quelltext mit Erklärung
/erklaere	Bittet um eine verständliche Erklärung
/zusammenfassen	Fasst einen längeren Text zusammen – höchstens fünf Punkte oder 200 Wörter
/plan	Erstellt eine nummerierte Schrittfolge

Befehl	Wirkung
/kritik	Prüft einen Text kritisch
/uebersetzen	Übersetzt
/brainstorm	Sammelt Ideen
/merken	Speichert etwas dauerhaft – ohne die KI zu fragen

> **Nichts kaputtzumachen**
> Ein unbekannter Schnellbefehl wird einfach als normale Frage behandelt.

Die Befehlspalette

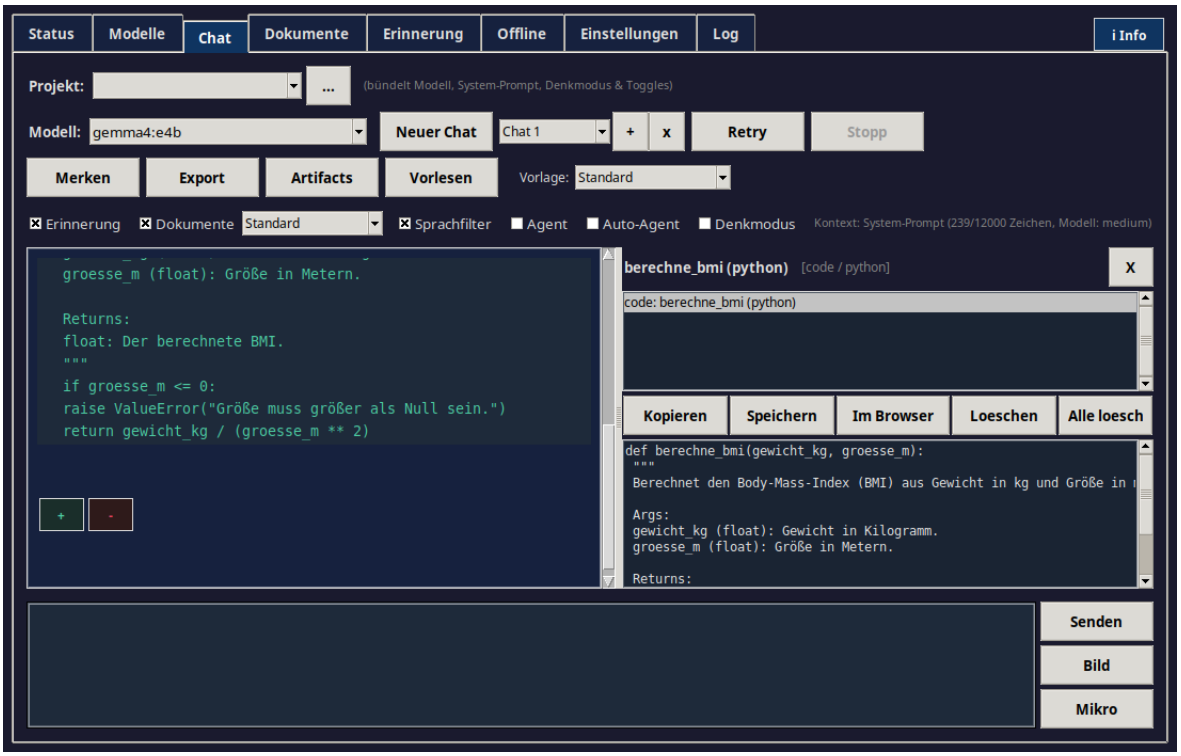
Strg + K öffnet ein Suchfeld. Tippen Sie ein paar Buchstaben, und es zeigt passende Projekte, Modelle, Schnellbefehle und Tabs. Mit den Pfeiltasten wählen, mit Eingabe ausführen, mit Escape schließen. Der schnellste Weg, wenn Sie wissen, was Sie wollen.



Die Befehlspalette.

Artefakte

Liefert die KI Quelltext oder eine HTML-Seite, erkennt das Programm das und legt es als Artefakt rechts neben dem Gespräch ab. 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. Für einen zweiten Bildschirm gibt es das Ganze auch als eigenes Fenster.



Links das Gespräch, rechts das erkannte Artefakt – die geteilte Ansicht.

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 – knapp, Code zuerst, großer Kontext, Auto-Agent an
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.

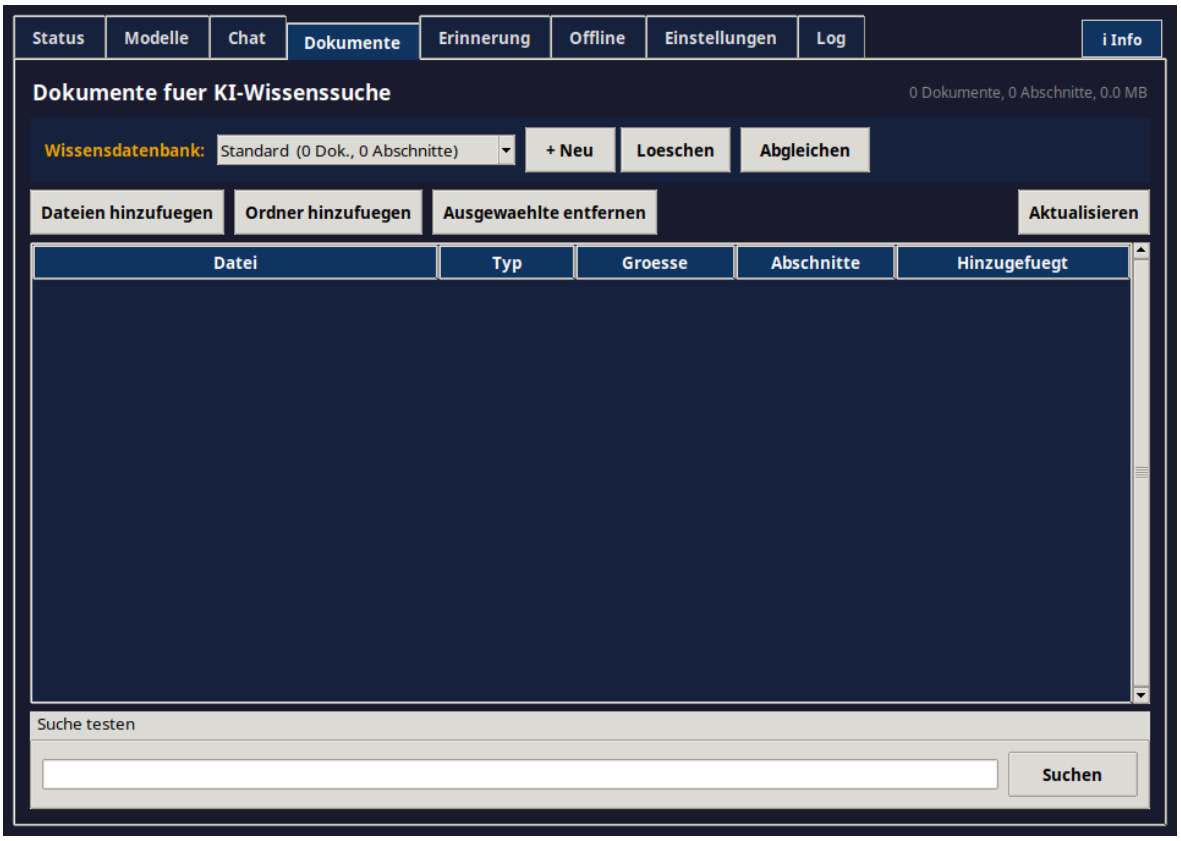
Beispiel für einen System-Prompt:

```
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.
```

> **Der letzte Satz ist der wichtigste**
> Ohne ihn erfinden kleine Modelle notfalls 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.

> **Ohne Einbettungsmodell keine Wissensdatenbank**
> Dafür wird das Modell `nomic-embed-text` (274 MB) gebraucht. Das Installationsprogramm bietet den Download an, sobald Ollama läuft; das Programm fragt außerdem nach, wenn Sie das erste Dokument hinzufügen. Fehlt das Modell, endet jeder Upload mit „Embedding fehlgeschlagen“. Nachholen im Terminal: `ollama pull nomic-embed-text`.

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

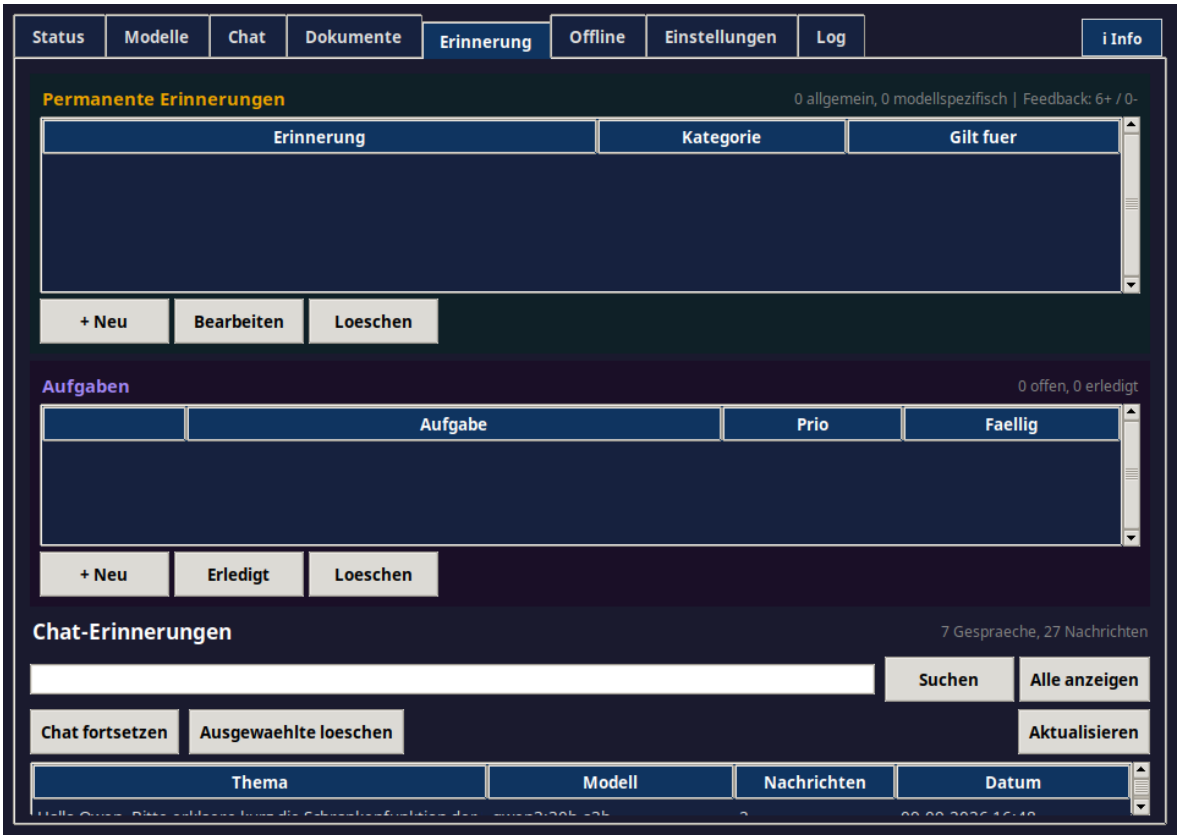
Der Knopf **Abgleichen** prüft alle Dokumente einer Datenbank: Geänderte Dateien werden neu eingelesen, gelöschte aus dem Bestand entfernt. Erkannt wird das an einer Prüfsumme, nicht am Datum – eine nur geöffnete Datei wird also nicht unnötig neu verarbeitet.

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.

> **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



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.

> **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 – auch nach Fragen mit Fragezeichen oder Sonderzeichen. 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. Auch hier laufen die Antworten Wort für Wort ein – zwischen den Werkzeugaufrufen und bei der Abschlussantwort.

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; kann pauschal freigegeben werden.	Suchen, Befehl ausführen, Code ansehen, Webseite abrufen, Web-Suche
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 verketten 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. Mit einer Beispielfunktion geprüft – entscheidend ist, ob das Modell einen strukturierten Aufruf zurückgibt statt Fließtext:

Modell	Antwort auf den Werkzeugtest	Agent-tauglich
qwen3:30b-a3b	echter Werkzeugaufruf	ja
gemma4:e4b	echter Werkzeugaufruf	ja
qwen2.5:3b	echter Werkzeugaufruf	ja
qwen2.5-coder:7b	gibt den Aufruf nur als Text aus	nein

Ein ungeeignetes Modell führt nicht zum Absturz – es beantwortet die Frage einfach normal, und das Programm 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.

> **Für Fortgeschrittene**
> Für den normalen Betrieb brauchen Sie diese Funktion 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-filesystem /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
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.

> **Nur freigeben, was Sie brauchen**
> Nutzen Sie den Knopf **Werkzeuge** und geben Sie nur frei, was Sie wirklich brauchen. Die Einstellungen zeigen Ihnen die Summe an und warnen ab 4.000 Zeichen.

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/deloma/voices/`.

Eine deutsche Stimme holen Sie sich so:

```
mkdir -p ~/.local/share/deloma/voices
cd ~/.local/share/deloma/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
```

> **Eine englische Stimme kann keine deutschen Laute bilden**
> Umlaute gehen verloren, das Ergebnis klingt falsch. Achten Sie darauf, dass der Dateiname mit `de_` beginnt. Beide Dateien – `.onnx` und `.onnx.json` – gehören zusammen.

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.

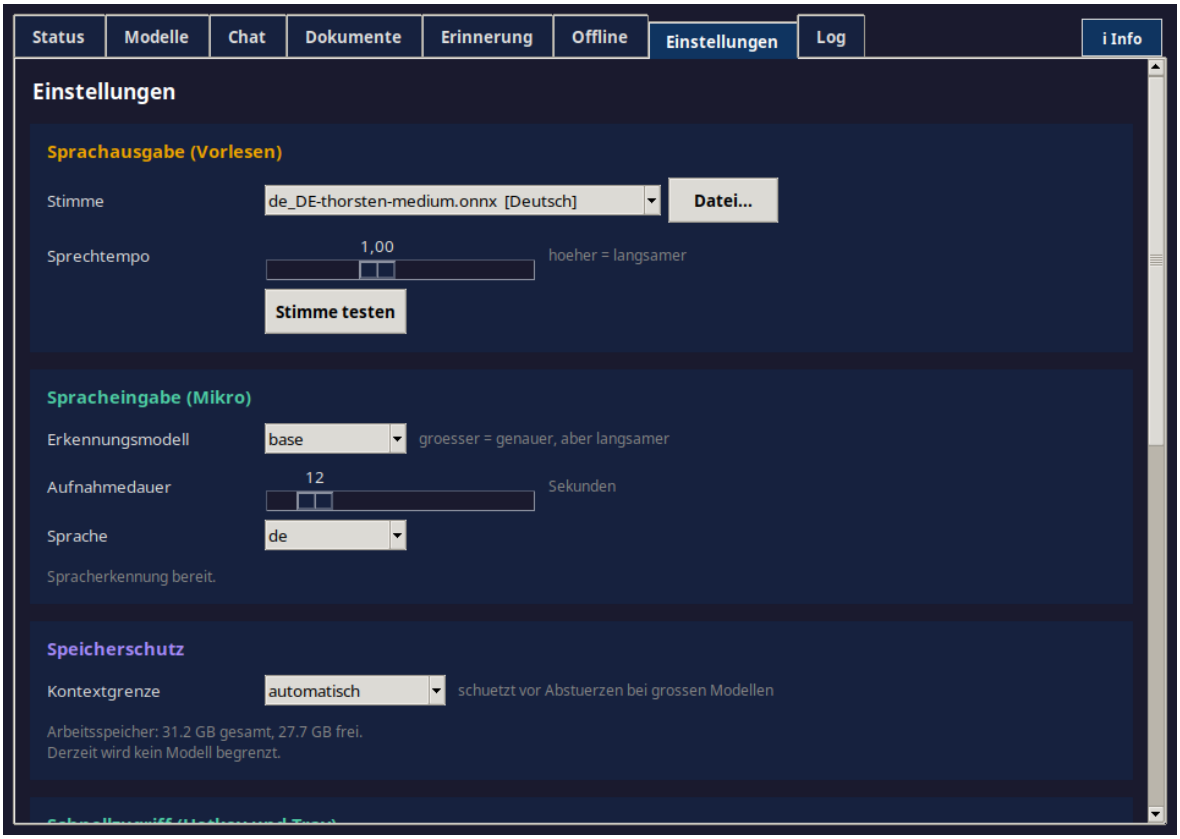
1. Auf **Mikro** klicken.
2. Ein Fenster zeigt einen Countdown. Sprechen Sie.
3. Nach Ablauf wechselt die Anzeige auf Verarbeite Aufnahme.
4. Der erkannte Text erscheint im Eingabefeld – Sie können ihn vor dem Absenden noch ändern.

> **Der erste Klick dauert länger**
> Beim ersten Mal nach dem Programmstart dauert es etwa zehn Sekunden länger: Das Erkennungsmodell wird in den Speicher geladen. Beim allerersten Mal überhaupt wird es zuvor einmalig heruntergeladen (rund 140 MB). 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

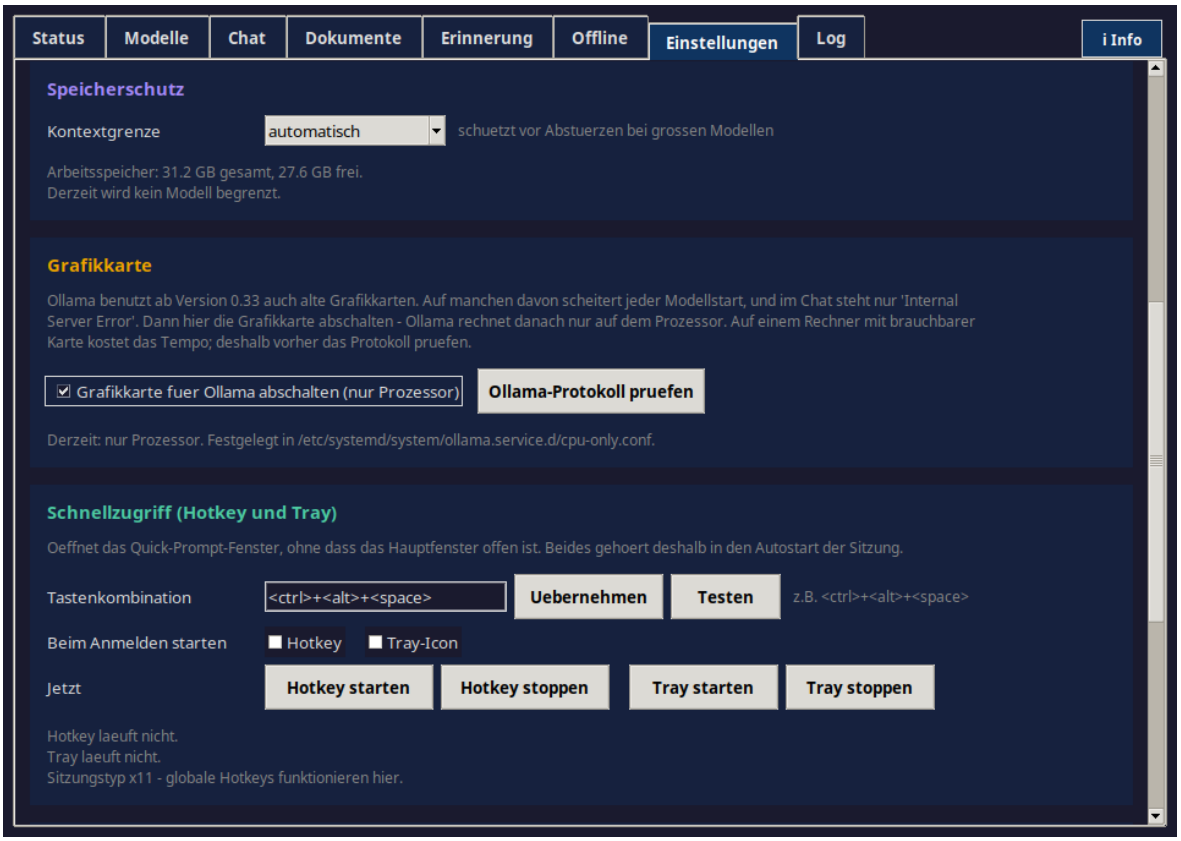
11. Einstellungen



Sprachausgabe, Spracheingabe und Speicherschutz.

Weiter unten auf derselben Seite: Grafikkarte, Schnellzugriff (Tastenkombination, Tray-Symbol, Autostart) und die MCP-Server.

Grafikkarte



Der Abschnitt Grafikkarte in den Einstellungen.

Ollama benutzt ab Version 0.33 auch ältere Grafikkarten. Auf manchen davon – etwa auf alten NVIDIA-Karten über den offenen Treiber – scheitert dann jeder Modellstart, und im Chat steht nur Internal Server Error. Der Haken **Grafikkarte für Ollama abschalten** zwingt Ollama auf den Prozessor. Dazu wird eine Datei unter `/etc/systemd/system/ollama.service.d/` angelegt und der Ollama-Dienst neu gestartet; dafür fragt das System einmal nach Ihrem Passwort.

Der Knopf **Ollama-Protokoll prüfen** liest das Protokoll des Ollama-Dienstes und sagt Ihnen, ob die Grafikkarte Ärger macht, ob sie sauber arbeitet oder ob Ollama ohnehin auf dem Prozessor rechnet. Prüfen Sie das, bevor Sie den Haken setzen.

> **Nur bei Bedarf**
> Auf einem Rechner mit brauchbarer Grafikkarte kostet der Haken Tempo, weil die Beschleunigung wegfällt. Setzen Sie ihn nur, wenn das Protokoll Fehler zeigt oder jedes Modell mit Internal Server Error scheitert. Der Haken lässt sich jederzeit wieder entfernen; auch das startet Ollama neu.

> **Von Hand angelegte Einstellung**
> Haben Sie die Grafikkarte früher selbst per Datei abgeschaltet, erkennt das Programm das und zeigt den Haken gesetzt. Beim Entfernen werden aus fremden Dateien nur die betreffenden Zeilen gelöscht; andere Einstellungen darin bleiben erhalten.

Alle Änderungen werden sofort gespeichert. Es gibt bewusst keinen Speichern-Knopf, den man vergessen kann. Die Einstellungen liegen in einer Datei:

```
~/.local/share/deloma/data/settings.json
```

12. Schnellzugriff

Beide Funktionen dieses Kapitels öffnen das Quick-Prompt – ein kleines Fenster für eine schnelle Frage, ohne das ganze Programm 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
```

> **Der Knopf Testen**
> Er 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**, **DeLoMa 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 geöffnetes Hauptfenster. 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
Stimmen	voices/	Stimmdateien für das Vorlesen
KI-Modelle	~/.ollama/models/ oder /usr/share/ollama/.ollama/models/	die Modelle selbst

Alles unterhalb von `~/.local/share/deloma/`, sofern nicht anders angegeben.

Sicherung

Beenden Sie das Programm und sichern Sie den Datenordner:

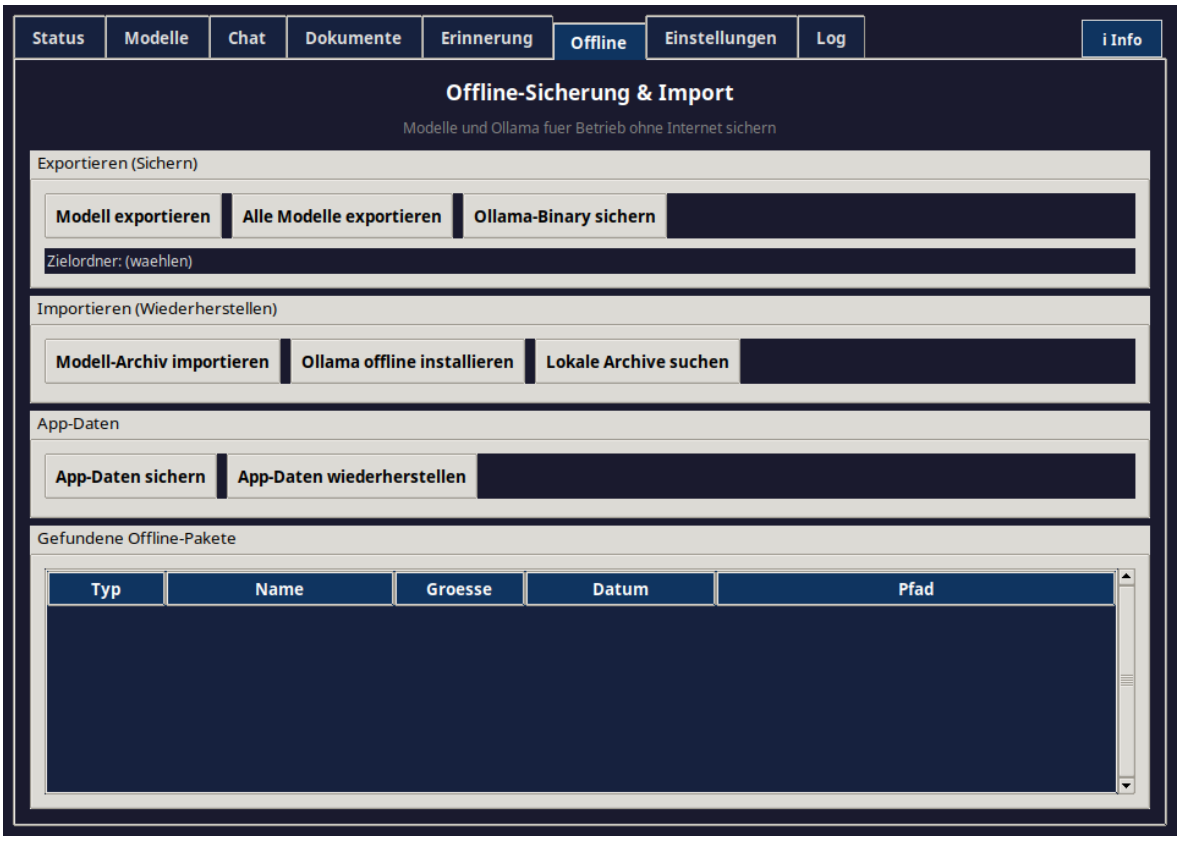
```
cp -r ~/.local/share/deloma/data ~/sicherung-$(date +%F)
```

Oder im Tab **Offline** mit **Programmdaten sichern** – das packt Gespräche, Erinnerungen und Wissensdatenbanken zusammen in ein Archiv.

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

14. Offline-Betrieb

Für Rechner ohne Internetanschluss – in Kliniken und Behörden eher Regel als Ausnahme. Der Tab **Offline** überträgt einen vollständig eingerichteten Arbeitsplatz auf ein Gerät, das nie online war.



Der Tab Offline.

Übertragen

- **Modelle exportieren** packt ein installiertes Modell als Archiv; **importieren** spielt es auf dem Zielrechner ein.
- **Ollama exportieren** sichert die Programmdatei von Ollama, damit sie auf dem Zielrechner aus der lokalen Kopie installiert werden kann.

Sichern

- **Programmdaten sichern** und **zurückspielen** – Gespräche, Erinnerungen und Wissensdatenbanken zusammen.
- **Archive suchen** findet vorhandene Pakete, um sie wiederzuverwenden.

Wo Ollama seine Modelle ablegt
Wurde Ollama als Dienst installiert (der Normalfall), liegen die Modelle unter `/usr/share/ollama/.ollama/models` und gehören dem Nutzer `ollama`. Das Programm findet dieses Verzeichnis von selbst. Lesen darf es aber nur, wenn Ihr Benutzerkonto in der Gruppe `ollama` ist – einmalig `sudo usermod -aG ollama $USER`, danach neu anmelden. Zeigt der Tab 0 Bytes, fehlt meist genau das. Der Import in ein Dienstverzeichnis bleibt Administratoren vorbehalten.

15. 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:

```
deloma
```

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.
command not found: deloma	`~/.local/bin` ist nicht im Suchpfad. Einmal ab- und wieder anmelden, oder `~/.local/bin/deloma` aufrufen.

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.

Jedes Modell scheitert mit „Internal Server Error“

Erscheint im Chat `[Agent-Fehler: Ollama-Verbindung: Internal Server Error]` bei jedem Modell, liegt es fast nie am Programm, sondern an Ollama und der Grafikkarte. Nachsehen:

```
journalctl -u ollama -b | grep -iE 'inference compute|Unsupported device|segmentation'
```

Steht dort Unsupported device oder ein Absturz von `llama-server`, kann Ollama die Grafikkarte nicht benutzen (typisch bei alten Karten über den offenen Treiber). Dann unter **Einstellungen → Grafikkarte** den Haken setzen – siehe Kapitel 11. Dasselbe von Hand im Terminal:

```
sudo mkdir -p /etc/systemd/system/ollama.service.d
printf '[Service]\nEnvironment="OLLAMA_VULKAN=false"\n' \
  | sudo tee /etc/systemd/system/ollama.service.d/cpu-only.conf
sudo systemctl daemon-reload && sudo systemctl restart ollama
```

> **Nur bei diesen Symptomen**
> Auf einem Rechner mit brauchbarer Grafikkarte schaltet dieser Eingriff die Beschleunigung ab. Nur anwenden, wenn das Protokoll die genannten Meldungen zeigt. Der Knopf **Ollama-Protokoll prüfen** in den Einstellungen nimmt Ihnen das Nachsehen ab.

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. Ist `nomic-embed-text` installiert?
Vergisst den Gesprächsverlauf	Kontext ist voll. Neuen Chat beginnen oder Erinnerungen ausdünnen

Alles ist im Akkubetrieb langsam

Manche Notebooks – gemessen am Fujitsu Celsius H730 – takten ohne Netzteil den Prozessor auf ein Fünftel herunter, egal was die Energieeinstellungen sagen. Die KI wird dann zwei- bis dreimal langsamer und die Antwortzeit schwankt stark. Prüfen:

```
cat /sys/class/power_supply/AC/online   # muss 1 ergeben
```

Für ernsthaftes Arbeiten mit der KI: Netzteil anschließen.

Der Auto-Agent ruft keine Werkzeuge auf

Fast immer liegt es am Modell. Nehmen Sie `qwen2.5:3b`, `gemma4:e4b` oder `qwen3:30b-a3b`. 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

```
ls ~/.local/share/deloma/venv/bin/piper
ls ~/.local/share/deloma/voices/
```

Fehlt das Programm oder die Stimmdatei, holen Sie beides nach – siehe Kapitel 10. Unter **Einstellungen → Sprachausgabe** steht im Klartext, was fehlt. Starten Sie das Programm über das Startmenü oder mit `deloma`; nur so ist das Sprachprogramm im Suchpfad.

Der Dokumente-Tab bleibt leer

Entweder fehlt das Paket `chromadb` in der Python-Umgebung, oder das Einbettungsmodell. Beides nachholen:

```
~/.local/share/deloma/venv/bin/python -m pip install -r ~/.local/share/deloma/requirements.txt
ollama pull nomic-embed-text
```

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/deloma
venv/bin/python -m unittest tests
```

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

16. Lizenz und Weitergabe

DeLoMa ist freie Software. Es 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**
> 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.

Der Name gehört nicht dazu

„DeLoMa – Deine Lokale Maschine“ ist die Bezeichnung dieses Projekts und nicht Teil der Lizenz. Veränderte Fassungen sind unter eigener Bezeichnung oder mit deutlich sichtbarem Zusatz zu verbreiten; die sachliche Nennung der Herkunft bleibt erlaubt. „Ollama“ ist eine Marke von Ollama Inc.; DeLoMa steht in keiner Verbindung zu Ollama Inc.

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/deloma/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

17. 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, Sicherung
Einstellungen	Sprache, Speicherschutz, Grafikkarte, Schnellzugriff, MCP
Log	Technische Meldungen



Der Info-Dialog (Knopf i Info oben rechts): Version, Impressum, Lizenz.

Wichtige Verzeichnisse

Pfad	Inhalt
~/.local/share/deloma/	Programm und Python-Umgebung (venv/)
~/.local/share/deloma/data/	Ihre Daten

Pfad	Inhalt
~/.local/share/deloma/voices/	Stimmen
~/.ollama/models/ bzw. /usr/share/ollama/.ollama/models/	KI-Modelle
~/.config/autostart/	Autostart-Einträge
~/deloma-sicherung/	Sicherung, die das Update angelegt hat

Nützliche Befehle

Befehl	Wirkung
deloma	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
venv/bin/python -m unittest tests	Selbsttest (im Programmordner)

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/deloma -name __pycache__ -exec rm -rf {} +`
4. Installation wiederholen: `bash install.sh`
5. Selbsttest laufen lassen.

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