

Ollama App Cheat Sheet

Lokale KI endlich einfach - Komplette Anleitung



Was ist die Ollama App?

Die neue Ollama App macht es endlich **kinderleicht, starke KI-Modelle kostenlos und sicher auf deinem eigenen Rechner** zu nutzen. Keine Terminal-Befehle mehr, keine komplizierte Einrichtung - einfach downloaden und loslegen!



Hauptvorteile:

• **Kostenlose, quelloffene Modelle** lokal nutzen • **Vollständige Datenkontrolle** - nichts verlässt deinen Computer • **Keine Cloud-Abhängigkeit** nach dem Download • **Benutzerfreundliche GUI** statt Kommandozeile • **Drag-and-Drop Dateien** für Dokumentenanalyse

Website: ollama.com (für alle Betriebssysteme verfügbar)



System-Anforderungen

Hardware-Empfehlungen:

Minimum (kleine Modelle 1B-3B): • **8GB RAM** - spürbar langsam, aber möglich • **2-5GB Festplatte** pro Modell

Empfohlen (7B-13B Modelle): • **16GB RAM** - angenehme Nutzung • **5-10GB Festplatte** pro Modell

Optimal (30B-70B+ Modelle): • **32GB+ RAM** - für große Modelle und Kontextfenster • **20-50GB Festplatte** pro großes Modell

Betriebssysteme:



Windows - direkt unterstützt



macOS - direkt unterstützt



Linux - direkt unterstützt

Beliebte Modelle (Beispiel: DeepSeek R1)

DeepSeek R1 Varianten:

Modell	Größe	RAM-Empfehlung	Verwendung
deepseek-r1:1.5b	1.1GB	4-6GB	Kleinste Option, sehr schwache Hardware
deepseek-r1:7b	4.7GB	8-12GB	Einstieg, moderate Hardware
deepseek-r1:latest (8b)	5.2GB	8-16GB	Empfohlene Standard-Version
deepseek-r1:14b	9.0GB	16-24GB	Erweiterte Fähigkeiten
deepseek-r1:32b	20GB	32-48GB	Komplexe Aufgaben, starke Hardware
deepseek-r1:70b	43GB	64GB+	Höchste Qualität, sehr starke Hardware
deepseek-r1:671b	404GB	512GB+	Experimentell, extremste Hardware

Alle DeepSeek R1 Modelle: 128K Kontext-Tokens (671b: 160K)

Weitere beliebte Modelle:

• **Gemma3** - Googles quelloffene Modell • **Qwen3** - Alibabas leistungsstarkes Modell • **Llama** - Metas bewährte Modell-Familie



Einfache Nutzung

Modelle herunterladen und nutzen:

1. **Ollama App öffnen**
2. **Modell-Dropdown klicken** - zeigt heruntergeladene + Empfehlungen
3. **Gewünschtes Modell suchen** - auch noch nicht heruntergeladene
4. **Erste Nachricht schreiben** - Download startet automatisch
5. **Antwort abwarten** - Modell wird geladen und antwortet

Praktische Features:

Drag-and-Drop Dateien: • **Dokumente in Chat ziehen** - Modell analysiert Inhalte • **Fragen zu Dokumenten stellen** - Informationen extrahieren • **Beispiel:** Q&A-Dokument hochladen und nach Details fragen

Konversationsverwaltung: • **Neue Gespräche** jederzeit starten • **Rechtsklick auf Konversationen** - umbenennen oder löschen • **Verlauf bleibt erhalten** für spätere Nutzung

Kontextlänge anpassen: • **Standard:** 4.000 Tokens • **Erweitert:** Bis zu 128.000 Tokens in Einstellungen • **Vorteil:** Längere Gespräche und größere Dokumente • **Nachteil:** Höherer RAM-Verbrauch

Erweiterte Funktionen

Spezielle Modelltypen:

Reasoning-Modelle (DeepSeek R1): • **Denkprozess sichtbar** - UI zeigt Reasoning-Verlauf • **Erweiterbarer Denkprozess** während der Antwort • **Komplexe Problemlösung** durch mehrstufiges Denken

Multimodale Modelle (Gemma3 Vision): • **Bilder per Drag-and-Drop** in Chat ziehen • **Bildbeschreibungen** und -analysen • **Visuelle Fragen** direkt stellen

Offline-Nutzung:

- ✓ **Komplett offline** nach Modell-Download
- ✓ **Keine Cloud-Verbindung** nötig für Nutzung
- ✓ **Daten verlassen nie** deinen Computer
- ✓ **Internet nur für** Download und Updates

Custom Models & System Prompts

Was sind Custom Models?

Custom Models ermöglichen es dir, **bestehende Modelle mit eigenen Charaktereigenschaften** zu versehen. Du kannst einem Modell eine spezielle Persönlichkeit, Expertise oder Verhaltensweise geben - z.B. einen freundlichen Assistenten oder einen Fachexperten.

Aktueller Status:

- ✗ **Nicht in der App verfügbar** - UI unterstützt noch keine direkte Erstellung
- ✓ **Workaround über Terminal** - einfacher als es klingt
- ✓ **Anschließend in App nutzbar** - nahtlose Integration

Wichtige Terminal-Befehle:

Modell erstellen:

```
ollama create [modellname] -f ./modelfile
```

Modell löschen:

```
ollama rm [modellname]
```

Alle Modelle anzeigen:

```
ollama list
```

Modell herunterladen:

```
ollama pull [modellname]
```

Allgemeine Syntax für Custom Models:

1. Modelfile erstellen: • Textdatei namens `modelfile` (ohne Endung) in einem Ordner erstellen

2. Modelfile-Struktur:

```
FROM [basismodell]

SYSTEM """
[Dein System-Prompt hier]
"""
```

3. Custom Model kompilieren:

```
ollama create [dein-modellname] -f ./modelfile
```

4. Custom Model nutzen: • In der Ollama App über Dropdown-Menü auswählen • Oder über Terminal:

```
ollama run [dein-modellname]
```

5. Custom Model löschen (optional):

```
ollama rm [dein-modellname]
```

Praktisches Beispiel: Custom Model "Jack" erstellen

1. Modelfile erstellen:

- **Neuen Ordner** auf deinem Computer öffnen • **Textdatei** mit Namen `modelfile` erstellen (ohne Endung) • **Folgenden Inhalt** einfügen:

```
FROM gemma3:1b
```

```
SYSTEM """
```

```
Du bist Jack, ein hilfreicher, sachlicher KI-Assistent. Du beantwortest  
Fragen mit zuverlässigen Informationen, bist präzise, freundlich und  
erklärst komplexe Zusammenhänge klar und verständlich. Antworte nach  
Möglichkeit auf Deutsch, halte dich an Fakten und strukturierte Auskünfte.  
Versuche, Unklarheiten aktiv nachzufragen.  
"""
```

2. Modell erstellen:

- **Terminal/Eingabeaufforderung** in dem Ordner öffnen • **Folgenden Befehl** ausführen:

```
ollama create jack -f ./modelfile
```

- **Warten** bis Erstellung abgeschlossen

3. In Ollama App nutzen:

- **Ollama App** wie gewohnt öffnen • **Modell-Dropdown** öffnen • **Nach "jack" suchen** und auswählen • **Direkt chatten** - Jack antwortet gemäß System-Prompt

Anpassungen und Variationen:

System-Prompt ändern: • **Modelfile bearbeiten** - Text unter SYSTEM anpassen • **Create-Befehl** erneut ausführen • **Altes Modell löschen:** `ollama rm jack` (optional)

Verschiedene Basismodelle: • **FROM gemma3:1b** → **FROM deepseek-r1:7b** • **Beliebiges heruntergeladenes Modell** als Basis nutzen

Weitere Charaktere erstellen: • **Experte:** Fachspezifische Beratung für bestimmte Bereiche
• **Tutor:** Geduldiger Lehrer für komplexe Themen • **Assistent:** Spezialisiert auf bestimmte Arbeitsabläufe

Ollama vs. Cloud-KI

Wann Ollama nutzen:

- ✓ **Datenschutz höchste Priorität** - Daten bleiben lokal
- ✓ **Offline-Arbeit erforderlich** - Keine Internet-Abhängigkeit
- ✓ **Kostenlos langfristig** - Keine laufenden API-Kosten
- ✓ **Volle Kontrolle gewünscht** - Custom Models und System-Prompts
- ✓ **Gute Hardware vorhanden** - RAM und Speicherplatz verfügbar
- ✓ **Experimentierfreude** - Verschiedene Modelle testen

Wann Cloud-KI nutzen:

- ✓ **Sofort startklar** - Keine Installation oder Hardware-Anforderungen
- ✓ **Höchste Performance** - Neueste, größte Modelle verfügbar
- ✓ **Team-Kollaboration** - Mehrere Nutzer gleichzeitig
- ✓ **Skalierbarkeit** - Automatische Anpassung an Bedarf
- ✓ **Wartungsfrei** - Updates und Optimierungen automatisch
- ✓ **Umfangreiche Integrationen** - APIs und Webdienste

Realistische Limitierungen

Hardware-Abhängigkeiten:

- **Performance variiert stark** je nach lokaler Hardware • **Ohne starke GPU** können große Modelle langsam sein • **RAM-Begrenzung** limitiert Modellgröße und Kontextlänge • **Festplattenspeicher** für mehrere große Modelle erforderlich

Funktionsumfang:

- **Kleinere Community** - weniger Tutorials und Support • **Begrenzte Modellauswahl** - nur Open-Source, keine GPT-4/Gemini • **Keine nativen Team-Features** - primär für Einzelnutzer • **Setup-Komplexität** für Custom Models über Terminal

Skalierungs-Herausforderungen:







- **Keine Lastverteilung** für Mehrbenutzerbetrieb • **Begrenzte API-Integrationen** verglichen mit Cloud-Anbietern • **Manuelles Modell-Management** erforderlich • **Keine automatische Skalierung** bei hoher Last

UI-Einschränkungen (aktuell):






- **Keine Custom Models** direkt in der App erstellbar • **System-Prompts** nur über Terminal konfigurierbar • **Erweiterte Einstellungen** teilweise nur über Kommandozeile

Ideale Zielgruppen

Perfekt für:

-  **Entwickler** - Custom Models und lokale Experimente
-  **Datenschutz-Bewusste** - Sensible Daten bleiben lokal
-  **Home-Office Nutzer** - Offline-Arbeit ohne Cloud-Abhängigkeit
-  **Studenten/Forscher** - Kostenlose Nutzung für Projekte
-  **Power-User** - Kontrolle über KI-Setup und -Konfiguration
-  **Experimentierfreudige** - Verschiedene Modelle und Setups testen

Vorsicht bei:

-  **Unternehmen** - Team-Features noch begrenzt
-  **Produktions-Umgebungen** - Skalierung und Support limitiert
-  **Mobile Nutzer** - Keine Smartphone-Apps verfügbar
-  **Schwache Hardware** - Große Modelle nicht nutzbar
-  **Schnellstarter** - Setup-Aufwand für erweiterte Features

Ausblick & Fazit

Aktueller Stand:

Die Ollama App ist ein **vielversprechender Start** in Richtung benutzerfreundlicher lokaler KI. Sie macht lokale Modelle endlich so **zugänglich wie Cloud-Services**, ohne die Datenkontrolle aufzugeben.

Stärken:

- **Revolutionäre Einfachheit** für lokale KI-Nutzung
- **Solide Grundfunktionen** bereits verfügbar
- **Starke Datenschutz-Position** in KI-Landschaft
- **Kostenfreies Modell** langfristig attraktiv

Verbesserungspotenzial:

- **UI-Features** für Custom Models in Entwicklung
- **Team-Funktionen** werden vermutlich folgen
- **Performance-Optimierungen** für verschiedene Hardware
- **Erweiterte Integrationen** zu erwarten

Empfehlung:

Jetzt einsteigen wenn:

- Du lokale Datenkontrolle schätzt
- Hardware ausreichend vorhanden ist
- Experimentierfreude für neue Technologie da ist
- Terminal-Nutzung für Custom Models ok ist

Noch abwarten wenn:

- Team-Features zwingend erforderlich sind
- UI-basierte Custom Model Erstellung wichtig ist
- Hardware-Upgrade erst geplant ist

Die Ollama App macht lokale KI endlich massentauglich - perfekt für alle, die Kontrolle über ihre Daten behalten wollen! 🤖🔒