Lokale KI auf dem Mac - Setup Guide

Empfohlene Lösung: Ollama + Open WebUI

Warum diese Kombination?

• ■ **100% kostenlos** und Open Source
• ■ **Läuft komplett lokal** (keine Cloud, keine API-Kosten)
• ■ **Einfache Installation** (Homebrew)
• ■ **Kann lernen** (Fine-tuning, RAG, Context Memory)
• ■ **Kann aktiv eingreifen** (Agenten, Tools, Auto-Fixes)
• ■ **Schöne Web-UI** (wie ChatGPT, aber lokal)

Installation (5 Minuten)

Schritt 1: Ollama installieren

```
# Mit Homebrew (empfohlen)
brew install ollama

# Oder direkt von ollama.ai
# https://ollama.ai/download
```

Schritt 2: Ollama starten

```
# Starte den Ollama-Server
ollama serve
```

Schritt 3: Erstes Modell herunterladen

```
# In einem neuen Terminal:
# Kleines, schnelles Modell (empfohlen für den Start)
ollama pull llama3.2:1b

# Oder größeres Modell (besser, aber langsamer)
ollama pull llama3.2:3b

# Oder sehr gutes Modell (langsamer, braucht mehr RAM)
ollama pull llama3.1:8b
```

Schritt 4: Open WebUI installieren

```
# Mit Docker (einfachste Methode)
docker run -d -p 3000:8080 --add-host=host.docker.internal:host-gateway -v open-webui:/app/backend/da

# Oder mit Docker Compose (empfohlen)
mkdir ~/open-webui
cd ~/open-webui
cat > docker-compose.yml <<EOF
version: '3.8'
services:
  open-webui:
    image: ghcr.io/open-webui/open-webui:main
    container_name: open-webui
    ports:
      - "3000:8080"
```

```
    volumes:
      - open-webui:/app/backend/data
    environment:
      - OLLAMA_BASE_URL=http://host.docker.internal:11434
    extra_hosts:
      - "host.docker.internal:host-gateway"
    restart: always
volumes:
  open-webui:
EOF

docker-compose up -d
```

Schritt 5: Öffne im Browser

```
http://localhost:3000
```

Erstelle einen Account und starte!

Was du damit machen kannst

*1. **Code-Assistent** (wie Cursor, aber lokal)*

• Code schreiben und debuggen
• Erklären, was Code macht
• Code refactoren
• Tests schreiben

*2. **Lernender Assistent***

• **RAG (Retrieval Augmented Generation)**: Lade deine Dokumentation/Code hoch, KI lernt daraus
• **Context Memory**: Erinnert sich an frühere Gespräche
• **Fine-tuning**: Trainiere auf deinen spezifischen Code-Stil

*3. **Aktives Eingreifen***

• **Agenten**: Kann selbstständig Aufgaben erledigen
• **Tools**: Kann Dateien lesen/schreiben, Commands ausführen
• **Auto-Fixes**: Kann Probleme automatisch erkennen und beheben

*4. **Spezielle Aufgaben***

• **Dokumentation generieren**
• **Code-Reviews**
• **Bug-Fixes vorschlagen**
• **Architektur-Entscheidungen**

Erweiterte Features

RAG Setup (KI lernt aus deinen Dateien)

1. In Open WebUI: Settings → RAG
2. Upload deine Dokumentation/Code
3. KI kann jetzt Fragen zu deinen Dateien beantworten

Agenten erstellen

1. In Open WebUI: Create → Agent
2. Definiere Tools (z.B. "read file", "run command")
3. Agent kann jetzt selbstständig arbeiten

Fine-tuning (auf deinen Code-Stil trainieren)

```
# Erstelle Trainingsdaten aus deinem Code
# (Format: JSONL mit Prompt/Response Paaren)

# Fine-tune mit Ollama
ollama create my-custom-model -f Modelfile
```

Alternative Lösungen

Option 2: LM Studio (GUI, sehr einfach)

• Download: https://lmstudio.ai
• Einfache GUI, keine Terminal-Befehle
• Gute Modell-Auswahl

Option 3: Jan.ai (Desktop App)

• Download: https://jan.ai
• Native Mac App
• Ähnlich wie LM Studio

Option 4: LocalAI (Self-hosted)

• Für fortgeschrittene Nutzer
• Mehr Kontrolle, komplexer Setup

Empfohlene Modelle für Mac

Für M1/M2/M3 Mac (Apple Silicon):

• **llama3.2:1b** - Sehr schnell, gut für Tests
• **llama3.2:3b** - Gute Balance
• **llama3.1:8b** - Sehr gut, braucht ~8GB RAM
• **mistral:7b** - Sehr gut für Code

Für Intel Mac:

• **llama3.2:1b** - Schnellste Option
• **llama3.2:3b** - Gute Balance

Nützliche Commands

```
# Modelle auflisten
ollama list

# Modell löschen
ollama rm model-name

# Modell-Info anzeigen
ollama show model-name

# Chat im Terminal
ollama run llama3.2:3b
```

Kosten

100% kostenlos!
• Keine API-Kosten
• Keine Cloud-Gebühren
• Läuft komplett lokal
• Nur Stromverbrauch (minimal)

Nächste Schritte

1. **Installiere Ollama** (5 Min)
2. **Lade ein Modell** (5-10 Min je nach Größe)
3. **Installiere Open WebUI** (5 Min)
4. **Teste es aus!**

Viel Erfolg! ■