Disclaimer:

Dieses Dokument dient lediglich als Hilfe zur Nutzung der beschriebenen Software. Der Autor übernimmt keine Verantwortung für etwaige Schäden an Hardware oder Software, die durch die Anwendung der Informationen in diesem Dokument entstehen könnten.

Ollama auf Linux

Ollama ist ein Open-Source Tool, das genutzt wird für den Betrieb von großen Sprachmodellen auf einem lokalen System.

Zu finden ist Ollama unter: https://ollama.com
Dort finden sich auch viele trainierte Sprachmodelle um Sie in Ollama laufen zu lassen.

Installation:

Ollama wird über den Terminal installiert, dazu muss man:

curl -fsSL https://ollama.com/install.sh | sh

eingeben. Die Software installiert sich.

Herunterladen von Sprachmodellen:

Man kann sich ein Sprachmodell direkt auf der Seite ollama.com aussuchen und herunterladen. Dazu wählt man ein Modell. Auf der entsprechenden Seite wählt man.

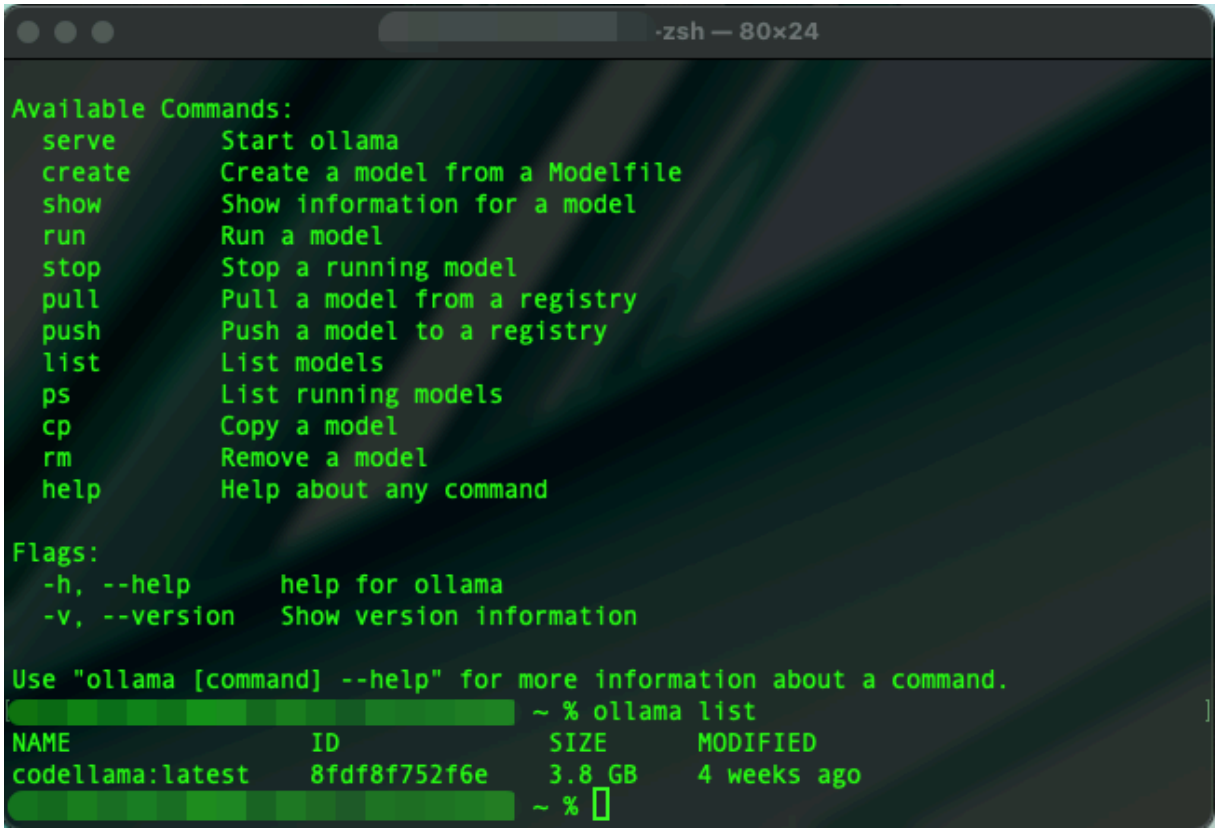
Im Bild haben wir ein Beispiel-Sprachmodell. Hier ist es starcoder2, ein Sprachmodell, das den Anwender beim Programmieren unterstützen soll.
Im Bild auf der linken Seite sieht man das ausgewählte Variante des Modells (3b). Dabei handelt es sich dann um ein Modell mit 3 Milliarden Trainingsdaten. Je höher dieser Wert ist umso eher kann man auch von brauchbaren Antworten aussehen. Aber Vorsicht, die Größe der Modelle steigt ebenfalls mit der Anzahl ihrer Trainingsdaten.

Im rechten Rahmen sehen wir den Befehl für die Kommandozeile, mit dem dieses Modell (starcoder2) installiert wird.

Nachdem das Modell heruntergeladen ist können wir es von der Kommandozeile aus starten.

In die Kommandozeile geben wir: *ollama*
Das Programm startet und zeigt eine Auswahl an Optionen, die uns bereitstehen.

Mit: *ollama list* können wir uns nun die Modelle anzeigen lassen, die wir heruntergeladen haben. Diese Liste könnte ungefähr so aussehen.



Im oberen Teil sieht man die Befehle, die man in Ollama absetzen kann.
Unten sieht man die Ausgabe von *ollama list*. Hier ist zu sehen, dass das Model codellama installiert ist.

Nun könnte man in unserem Beispiel seine KI lokal in Ollama starten.
Dazu gibt man: *ollama run codellama* ein. Das Model sollte starten und danach bereit für die Nutzung sein.