



Laptop
Endgerät

Umsetzung mit Lokaler KI

oer.one/ki-lokal/alf

Hardware Voraussetzung

- Technisches Know-How: mittel
- Betriebssystem: Windows, macOS oder Linux
- RAM: mindestens 8 GB
- Grafikkarte: keine Voraussetzung
- Speicherplatz: je nach Modell, ab 2 GB
- Internet: nach dem Download nicht notwendig

Anleitung

Schritt 1: Webseite öffnen

Rufe die Projektseite Llamafile auf (z. B. über [Mozilla.ai](#) oder [GitHub](#)) und wähle ein passendes Llamafile-Modell aus der Liste der Beispielmmodelle aus.

Schritt 2: Llamafile herunterladen

Lade die gewünschte Datei herunter. Es handelt sich um eine einzelne Datei, die bereits das Modell und die Chat-Oberfläche enthält.

Schritt 3: Datei ausführbar machen

- Windows: Datei in *.exe umbenennen
- macOS / Linux: Datei ausführbar machen (chmod +x dateiname)

Schritt 4: KI nutzen

Llamafile in der Kommandozeile starten.

Die KI läuft nun vollständig lokal auf dem eigenen Gerät. Fragen können direkt in der Kommandozeile oder im Browser gestellt werden. Es ist keine Anmeldung oder Cloudzugang erforderlich.

Model	Size	License	Llamafile	other quants
LLaMA 3.2 1B Instruct	1.11 GB	LLaMA 3.2	Llama-3.2-1B-Instruct.Q6_K.llamafile	See HF repo
LLaMA 3.2 3B Instruct	2.62 GB	LLaMA 3.2	Llama-3.2-3B-Instruct.Q6_K.llamafile	See HF repo
LLaMA 3.1 8B Instruct	5.23 GB	LLaMA 3.1	Llama-3.1-8B-Instruct.Q4_K_M.llamafile	See HF repo
Gemma 3 1B Instruct	1.32 GB	Gemma 3	gemma-3-1b-it.Q6_K.llamafile	See HF repo
Gemma 3 4B Instruct	3.50 GB	Gemma 3	gemma-3-4b-it.Q6_K.llamafile	See HF repo
Gemma 3 12B Instruct	7.61 GB	Gemma 3	gemma-3-12b-it.Q4_K_M.llamafile	See HF repo

Liste an Llamafiles

Quelle: https://mozilla-ai.github.io/llamafile/example_llamafiles



Lokale Chat-Oberfläche im Browser



CC BY 4.0 International

Benedikt Brünner (TU Graz), Andreas Zitek (BOKU University), Martin Ebner (TU Graz)