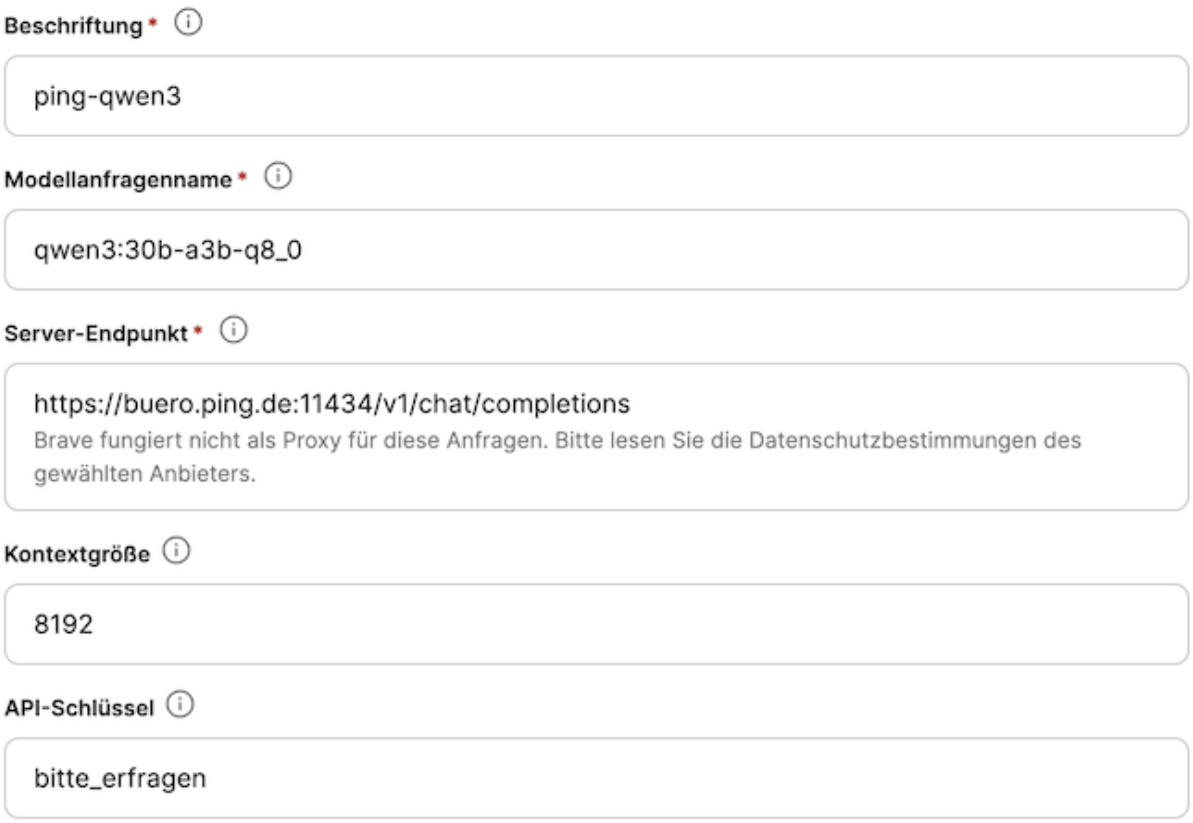
Brave-Leo-AI mit PING KI

Konfigurieren von Leo AI im Brave Browser mit dem PING AI Server

Auf cogito (KI-Server) läuft ein Ollama-server, der einen OpenAI-API kompatiblen Endpunkt bereitstellt. Er ist unter https://ki.ping.de:8000/ erreichbar, man benötigt für den Zugriff ein Bearer Token.

1. Besorgt euch das Bearer Token, siehe API-Token Seite (für Mitglieder)
2. Im Brave Browser oben rechts auf das "Leo AI" Icon  klicken.
3. Klickt oben rechts auf die 3 Punkte übereinander " ⋮ "
4. Klickt ganz unten auf "Erweiterte Einstellungen" (mit dem Zahnrad). Ihr landet dann auf `brave://settings/leo-ai`
5. Unter "Bringen Sie Ihr eigenes Modell mit" klickt auf "Neues Modell hinzufügen"
6. Macht folgende Einstellungen:
   - Beschriftung: `ping-qwen3`
   - Modellanfragename: `qwen38-27b` (diesen Namen seht ihr so auch im open-webui oder s.u.)
   - Server-Endpunkt: `https://ki.ping.de:8000/v1/chat/completions`
   - Kontext-Größe: `65536` (mehr geht auch, hängt vom LLM und freien VRAM ab).
   - API-Schlüssel: Siehe Punkt 1. Ohne "Bearer" davor eingeben.
     ⚠ ACHTUNG, Screenshot **veraltet**:



7. Klickt auf "Modell speichern"
8. Stellt das "Standardmodell für neue Unterhaltungen" auf `ping-qwen3`

Fertig. Wenn ihr jetzt auf das Leo-AI-Icon klickt startet eine neue Unterhaltung mit dem LLM auf dem PING Server cogito. Wenn ihr auf das "Seitenleiste anzeigen" Icon daneben klickt teilt sich das Browserfenster und ihr seht neben der Webseite das Leo AI Chatinterface, dort könnt ihr dann das LLM zur gerade aktiven Webseite befragen (zusammenfassen etc.).

Die KI von der Brave Search läuft davon unabhängig in der Cloud von Brave.

Verfügbare Modelle auflisten

Wenn ihr eine Liste aller installierten Modelle sehen möchtet, dann könnt ihr das entweder in open-webui oder es geht über die Ollama API wie folgt (ihr benötigt die Befehle `curl` und `jq`):

```
BEARER_TOKEN=siehe_oben
curl -sH "Authorization: Bearer $BEARER_TOKEN" https://ki.ping.de:8000/v1/models | jq
```

Revision #11
Created 2026-01-31 16:49:09 UTC by Daniel Hess
Updated 2026-08-30 09:11:23 UTC by Sven Neuhaus