VoiceMem

Ein Streaming-Speichersystem für Echtzeit-Sprachagenten mit getrennter Fakten- und Emotionsverarbeitung.

STERNE	LIZENZ	SPRACHE	VERSION	ANGELEGT	ZULETZT GEÄNDERT
1.240	Apache-2.0	Python	v0.0.2	2026-08-17	2026-09-05

Was es ist

VoiceMem ist ein Python-Framework, das Sprachassistenten ein Langzeitgedächtnis verleiht. Es spaltet die Speicherung in einen linken Teil für faktenbasierte Informationen und einen rechten Teil für Emotionen, Persönlichkeit und Beziehungen. Die Architektur arbeitet vollständig streaming-basiert, sodass die Speicherabfrage bereits während der Spracheingabe läuft.

Was es kann

- Fakten und Emotionen getrennt speichern und abrufen (Left Brain / Right Brain).
- Audio in Echtzeit verarbeiten mit einer Latenz von 134 ms bei der Abfrage.
- Während des Sprechens bereits relevante Erinnerungen vorausschauend laden (0–300 ms spekulatives Prefetching).
- Aus Sprache automatisch Entitäten, Schemata, Sprecher-IDs, Emotionen und Embeddings extrahieren.
- Nur die Top-5 relevantesten Erinnerungen in den Kontext injizieren (ca. 430 Token).
- Als Offline-Motor oder als integrierter Stream in eigene Agenten eingebunden werden.

So fängst du an

Die Installation setzt Python 3.10 oder höher voraus und erfordert das Herunterladen von Modellen über Hugging Face.

```
git clone https://github.com/xzf-thu/VoiceMem.git
cd VoiceMem
pip install voicemem
pip install "voicemem[slm]"
pip install -U huggingface_hub
hf download zhifeixie/VoiceMem_Default_Models_Env --local-dir ./models
```

Grenzen und Bedingungen

- Für Chat, TTS und Realtime-Funktionen wird ein OpenAI-API-Key benötigt; ohne Konto bei OpenAI sind diese Teile nicht nutzbar.
- Die interaktive Web-Demo (python web/run.py) liegt nur im Repository, nicht im pip-Paket.

- Die lokale Spracherkennung basiert auf chinesischsprachigen Modellen (paraformer-zh-streaming); die Unterstützung anderer Sprachen ist nicht dokumentiert.
- Die Version v0.0.2 ist erst wenige Wochen alt und hat 7 offene Issues; Produktivbetrieb erfordert eigene Tests.
- Eigenes Fine-Tuning erfordert ms-swift 4.5.2, bitsandbytes und eine GPU mit ausreichend VRAM (Details in finetune/README.md).

Für wen sich das lohnt

VoiceMem lohnt sich für Entwickler von Sprachagenten, die echtes Langzeitgedächtnis mit geringer Latenz brauchen und bereit sind, ein OpenAI-Konto zu verknüpfen. Wer rein lokale Systeme ohne externe APIs sucht oder stabile, langjährig erprobte Software benötigt, sollte abwarten oder auf reine Vektordatenbanken wie Mem0 setzen.

Zusammengestellt von **localdev.ai**
Quelle: **github.com/xzf-thu/VoiceMem** · Stand 2026-09-05

Grundlage: Repository-Metadaten: Sterne 1240, Sprache Python 84.8%, Lizenz Apache-2.0, Version v0.0.2, 7 offene Issues; Installationsabschnitt: Python 3.10+, git clone, pip install voicemem, pip install "voicemem[slm]", pip install -U huggingface_hub, hf download; Basis-Usage: VoiceMem(mode="normal", openai_key="api_xxx", top_k=5), vm.ingest(), vm.search(), vm.stream(); Architektur: Left Brain (Schema/Entity