Healthy AI First, Folge 10: Praxis 2, Mein Sprachassistent am Second Brain

Quellen und Belege zur Folge

Die zweite Praxisfolge zeigt Hall, meinen selbst gebauten Sprachassistenten für mein Second Brain: Architektur, Messwerte, der Vergleich lokal gegen Cloud und acht Bauregeln. Alle Messwerte stammen aus meinen eigenen Tests im November 2026 auf einem Mac mini mit 64 GB.

Alle Links wurden am 16.11.2026 geprüft.

Was wir in dieser Folge erwähnen

Warum selbst bauen

- **Aussage:** Mit ChatGPT kann man einfach reden; mein Setup läuft auf Claude, mit Claude kann ich so nicht sprechen, und in ChatGPTs Gesprächsfunktion bekomme ich mein Setup nicht hinein.
  Quelle: Eigene Erfahrung.
  Einordnung: Persönliche Einschätzung für mein Setup, kein Produktvergleich.
- **Aussage:** Hermes, ein Agenten-System, hat einen Sprachmodus; ich habe ihn getestet. Hermes wäre eine zweite Plattform neben meinem Brain gewesen, und die offizielle iPhone-App gab es nur über eine Warteliste.
  Quelle: Eigener Test, Hermes-Dokumentation.
  Link: https://hermes-agent.nousresearch.com/docs
  Einordnung: Stand Mitte November 2026; die Verfügbarkeit der App kann sich ändern.

Architektur

- **Aussage:** Turn-Modell, Whisper (große Version) lokal, ein Qwen-Modell mit 9 Milliarden Parametern über Ollama, gut dreißig Werkzeuge, Suche über rund 1.600 Notizen, Claude als Helfer, Qwen-TTS als Stimme.
  Quelle: Eigener Aufbau.
  Link: https://arxiv.org/abs/2212.04356 (Whisper), https://github.com/QwenLM/Qwen3-TTS, https://ollama.com
  Einordnung: Beschreibung meiner Einrichtung, keine Empfehlung für genau diese Modelle.
- **Aussage:** Eine Namensliste für die Spracherkennung hat in meinen Tests ungefähr so viel gebracht wie ein größeres Modell.
  Quelle: Eigener Test.
  Einordnung: Test mit künstlichen Stimmen und erzeugtem Lärm (Wortfehler ohne Liste 8,3 Prozent, mit Liste 6,6 Prozent, kleineres Whisper-Modell); kein Test mit echten Telefonaufnahmen.

Latenzbudget

- **Aussage:** Die Pausen zwischen zwei Sprechern liegen bei etwa zweihundert Millisekunden; wir planen die Antwort, während der andere noch spricht.
  Quelle: Levinson, S. C. und Torreira, F. (2015): Timing in turn-taking and its implications for

processing models of language. Frontiers in Psychology 6:731.
Link: https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00731/full
Einordnung: Übersichtsarbeit; über Sprachen hinweg ähnlich, siehe Stivers et al. (2009), PNAS, https://pubmed.ncbi.nlm.nih.gov/19553212/

- **Aussage:** Erste Version 19 bis 35 Sekunden pro Antwort; heute meist 1 bis 2,5 Sekunden nach erkanntem Satzende; erste Antwort nach dem Laden knapp 9 Sekunden; Ohren und Stimme rund 7 GB, Sprachmodell etwa 9 GB.
  Quelle: Eigene Messungen (31.10. bis 12.11.2026).
  Einordnung: Kleine Stichproben, ein Rechner.

Turn-Erkennung und Echo

- **Aussage:** Smart Turn von Pipecat (etwa 8 Millionen Parameter) entscheidet bei Pausen, ob ein Satz fertig ist, und wartet bei unfertigen Sätzen bis zu 3 Sekunden.
  Quelle: Pipecat Smart Turn, eigener Aufbau.
  Link: https://github.com/pipecat-ai/smart-turn
  Einordnung: Eigene Schwellen, im Alltag weiter beobachtet.

Unterschrift statt Okay

- **Aussage:** Prompt Injection ist das größte Risiko für Anwendungen mit Sprachmodellen.
  Quelle: OWASP Top 10 for LLM Applications 2025, LLM01.
  Link: https://genai.owasp.org/llm-top-10/
  Einordnung: Brancheneinschätzung; unsere Regel "vorgelesen heißt nicht beauftragt" ist ein Prinzip, kein vollständiger Schutz.

Testen mit echten Sätzen

- **Aussage:** 84 Prüfszenarien, 139 echte Sätze und 16 Gespräche laufen vor jedem Neustart.
  Quelle: Eigener Aufbau.
  Einordnung: Die Kriterien habe ich selbst gesetzt; Läufe schwanken um einige Punkte.

Lokal gegen Cloud

- **Aussage:** Auf 84 Szenarien bestanden lokales Qwen 76, Claude Sonnet 75, Claude Haiku (schnelle Variante) 66.
  Quelle: Eigene Messung (12.11.2026).
  Einordnung: Einzelne Läufe, Schwankung etwa plus minus zwei.
- **Aussage:** Gemini Live als Speech-to-Speech: im Mittel rund 1,5 Sekunden Antwortzeit, etwa 11 bis 12 Cent pro Minute; im ersten Test erfand das Modell Entscheidungen.
  Quelle: Eigene Tests (15. und 16.11.2026), Gemini Live API.
  Link: https://ai.google.dev/gemini-api/docs/live-api
  Einordnung: Kurzer Test mit neun Messungen; die Messung beginnt anders als bei der lokalen Kette, der Vergleich ist daher nur grob.

Healthy AI First ist KI-recherchiert und KI-generiert, gesprochen von einer KI-Version meiner Stimme und von mir (Björn) geprüft.