Healthy AI First, Episode 10: Practice 2, My Voice Assistant for My Second Brain

Sources and evidence for this episode

The second practice episode shows Hall, my self-built voice assistant for my second brain: architecture, measurements, the local vs cloud comparison and eight build rules. All measurements come from my own tests in November 2026 on a Mac mini with 64 GB.

All links checked on 16 Nov 2026.

What we mention in this episode

Why build your own

- **Claim:** With ChatGPT you can just talk; my setup runs on Claude, I can't talk to Claude that way, and I can't plug my setup into ChatGPT's voice feature.
  Source: Own experience.
  Context: Personal assessment for my setup, not a product comparison.
- **Claim:** Hermes, an agent system, has a voice mode; I tested it. Hermes would have been a second platform next to my brain, and the official iPhone app was only available through a waitlist.
  Source: Own test, Hermes documentation.
  Link: https://hermes-agent.nousresearch.com/docs
  Context: As of mid-November 2026; app availability may change.

Architecture

- **Claim:** Turn model, Whisper (large version) locally, a Qwen model with 9 billion parameters on Ollama, a little over thirty tools, search across about 1,600 notes, Claude as a helper, Qwen TTS as the voice.
  Source: Own setup.
  Link: https://arxiv.org/abs/2212.04356 (Whisper), https://github.com/QwenLM/Qwen3-TTS, https://ollama.com
  Context: Description of my setup, not a recommendation for these exact models.
- **Claim:** A name list for speech recognition helped about as much as a bigger model in my tests.
  Source: Own test.
  Context: Test with synthetic voices and generated noise (word error rate 8.3 percent without, 6.6 percent with the list, smaller Whisper model); not tested on real phone recordings.

Latency budget

- **Claim:** Gaps between speakers are about two hundred milliseconds; we plan our answer while the other person is still talking.
  Source: Levinson, S. C. and 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

Context: Review article; similar across languages, see Stivers et al. (2009), PNAS, https://pubmed.ncbi.nlm.nih.gov/19553212/

- **Claim:** First version 19 to 35 seconds per answer; today usually 1 to 2.5 seconds after the end of the sentence is detected; first answer after loading almost 9 seconds; ears and voice about 7 GB, language model about 9 GB.
  Source: Own measurements (31 Oct to 12 Nov 2026).
  Context: Small samples, one computer.

Turn detection and echo

- **Claim:** Smart Turn by Pipecat (about 8 million parameters) decides at pauses whether a sentence is finished and waits up to 3 seconds when it sounds unfinished.
  Source: Pipecat Smart Turn, own setup.
  Link: https://github.com/pipecat-ai/smart-turn
  Context: Own thresholds, still being observed in daily use.

Signature, not okay

- **Claim:** Prompt injection is the biggest risk for applications built on language models.
  Source: OWASP Top 10 for LLM Applications 2025, LLM01.
  Link: https://genai.owasp.org/llm-top-10/
  Context: Industry assessment; our rule "read aloud is not an order" is a principle, not complete protection.

Testing with real sentences

- **Claim:** 84 test scenarios, 139 real sentences and 16 conversations run before every restart.
  Source: Own setup.
  Context: I set the criteria myself; runs vary by a few points.

Local vs cloud

- **Claim:** On 84 scenarios, local Qwen passed 76, Claude Sonnet 75, Claude Haiku (fast variant) 66.
  Source: Own measurement (12 Nov 2026).
  Context: Single runs, variation about plus or minus two.
- **Claim:** Gemini Live as speech-to-speech: about 1.5 seconds average response time, about 11 to 12 euro cents per minute; in the first test the model made up decisions.
  Source: Own tests (15 and 16 Nov 2026), Gemini Live API.
  Link: https://ai.google.dev/gemini-api/docs/live-api
  Context: Short test with nine measurements; measured from a different starting point than the local chain, so the comparison is rough.

Healthy AI First is AI-researched and AI-generated, narrated in an AI version of my voice and reviewed by me (Björn).