

Where it breaks before it touches a patient or a student.

Consolidate, don't expand · squeeze what already exists

Updated October 7, 2026

AGENTS

- 1:1 support agent** ↗
WhatsApp · answers health-companion users, opens tickets, learns from every conversation.
- Group agent** ↗
WhatsApp · its own identity and model per group, keeps context, encourages participation.
- Collections agent**
Telegram · payment follow-up wired to the CRM.
- Multimodal agent**
Voice, text and image, over Telegram and web.
- Prototyping bench**
Where new things are tried before they reach production; currently home to the multimodal agent.
- Personal assistant** ↗
Telegram and web · long-term memory, email digests, reminders and tasks.
Raspberry Pi 5 · 16 GB — Mac mini M1 · 8 GB (prototyping bench)

BRAIN / ORCHESTRATION

- Orchestration station**
It coordinates; it doesn't do the heavy lifting.
 - Coding-agent orchestrator**
Claude Code, with a second opinion from Codex and Antigravity CLI.
 - Local speech and text models**
MLX.
 - Secrets and single sign-on**
Encrypted vault with rotation + single sign-on · Vault + Keycloak.
- Mac mini M4 Pro · 64 GB

SERVICES

- Internal clinical hub**
Cancer center · internal tools, access per user.
- SimAcademy** ↗
Moodle-based medical education platform · courses and certificates.
- Asesores en Emergencias**
Clinical training platform.
- AI-assisted marketing studio** ↗ **IN PRODUCTION**
An agent with its own interface: images, brand voice and publishing for the in-house brands. Runs on the same machine as the clinical training platform.
- SimStage** ↗
Live class production.
- Health Companion** ↗
Health companion, in the cloud.
- Browser arm (Explorer)**
Shared by the orchestrator and the personal assistant.

LLM router

One door to every model; switching models doesn't mean rewriting agents.

INFERENCE

- Local inference with vision** ↗
A 35B MoE Qwen model that reads images, served locally with vLLM, at no cost per use.
 - Irreversible-action classifier** **IN PRODUCTION**
Stops destructive actions before they run.
 - Cloud models**
Anthropic, OpenAI, Gemini and others, always through the router.
 - Model evaluation arenas** ↗
I pit models against each other on my own cases and criteria — task success, tool-use safety, calibration, latency— before anything reaches production.
- AI workstation · 128 GB unified memory

STORAGE AND BACKUP

- Primary storage**
9 TB · fleet library and private file cloud (Nextcloud).
- Offsite backup**
Encrypted copy off site · borg · 3-2-1 scheme.

NETWORK

- Encrypted private network**
Across every node · Tailscale.
- Three paths to the internet**
Fiber · cellular SIM · satellite, with automatic failover.

MONITORING

- Per-service watchdogs that restart and report.
- Twice-weekly fleet review: it detects and proposes; nothing runs without authorization.
- Phone alerts through the personal assistant.

GROUND RULES

- Consolidate, don't expand**
Squeeze what exists before adding hardware or complexity.
- A gate for the irreversible**
Destructive actions are stopped in code, not in the prompt.
- Secrets only in the vault**
Never in repos or loose files; rotated and distributed.
- Nothing without written authorization**
Anything touching production is approved first.