

Donde se rompe antes de tocar a un paciente o a un alumno.

Consolidar, no expandir · exprimir lo que ya existe

Actualizado 7 de octubre de 2026

AGENTES

- Agente de soporte 1:1** ↗
WhatsApp · atiende a usuarios del acompañante de salud, levanta tickets y aprende de cada conversación.
- Agente de grupos** ↗
WhatsApp · identidad y modelo propios por grupo, mantiene contexto y promueve participación.
- Agente de cobranza**
Telegram · seguimiento de pagos conectado al CRM.
- Agente multimodal**
Voz, texto e imagen, por Telegram y por web.
- Banco de prototipos**
Donde se prueba lo nuevo antes de que llegue a producción; hoy sostiene al agente multimodal.
- Asistente personal** ↗
Telegram y web · memoria longitudinal, resúmenes de correo, recordatorios y tareas.
Raspberry Pi 5 · 16 GB — Mac mini M1 · 8 GB (banco de prototipos)

CEREBRO / ORQUESTACIÓN

- Estación de orquestación**
Coordina; no carga lo pesado.
 - Orquestador de agentes de código**
Claude Code, con segunda opinión de Codex y Antigravity CLI.
 - Modelos locales de voz y texto**
MLX.
 - Secretos y acceso único**
Bóveda cifrada con rotación + inicio de sesión único · Vault + Keycloak.
- Mac mini M4 Pro · 64 GB

SERVICIOS

- Hub clínico interno**
Centro oncológico · herramientas internas por usuario.
- SimAcademy** ↗
Plataforma de educación médica sobre Moodle · cursos y certificados.
- Asesores en Emergencias**
Plataforma de formación clínica.
- Estudio de marketing asistido por IA** ↗ **EN PRODUCCIÓN**
Agente con interfaz propia: imágenes, voz de marca y publicación para las marcas propias. Corre en la misma máquina que la plataforma de formación clínica.
- SimStage** ↗
Producción de clases en vivo.
- Health Companion** ↗
Acompañante de salud, en la nube.
- Brazo navegador (Explorador)**
Compartido por el orquestador y el asistente personal.

Router de LLMs

Puerta única a todos los modelos; cambiar de modelo no obliga a reescribir agentes.

INFERENCIA

- Inferencia local con visión** ↗
Modelo Qwen de 35B MoE que lee imágenes, servido localmente con vLLM, sin costo por uso.
 - Clasificador de acciones irreversibles** **EN PRODUCCIÓN**
Detiene lo destructivo antes de ejecutarlo.
 - Modelos en la nube**
Anthropic, OpenAI, Gemini y otros, siempre a través del router.
 - Arenas de evaluación de modelos** ↗
Comparo modelos entre sí con mis propios casos y criterios —éxito en la tarea, seguridad con herramientas, calibración, latencia— antes de que algo llegue a producción.
- Estación de IA · 128 GB de memoria unificada

ALMACENAMIENTO Y RESPALDO

- Almacenamiento primario**
9 TB · biblioteca de la flota y nube privada de archivos (Nextcloud).
- Respaldo offsite**
Copia cifrada fuera del sitio · borg · esquema 3-2-1.

RED

- Red privada cifrada**
Entre todos los nodos · Tailscale.
- Tres salidas a internet**
Fibra óptica · SIM celular · satélite, con conmutación automática.

VIGILANCIA

- Watchdogs por servicio que reinician y avisan.
- Revisión de flota dos veces por semana: detecta y propone; nada se ejecuta sin autorización.
- Avisos al teléfono a través del asistente personal.

REGLAS DE ORO

- Consolidar, no expandir**
Exprimir lo que ya existe antes de agregar equipo o complejidad.
- Compuerta para lo irreversible**
Las acciones destructivas se frenan en código, no en el prompt.
- Secretos solo en la bóveda**
Nunca en repositorios ni archivos sueltos; se rotan y se distribuyen.
- Nada sin autorización escrita**
Lo que toca producción se aprueba primero.