REF: DOC-CIW-2026-AI
EDICIÓN: Rev. 2026.2 (Oficial)
ÁREA: Inteligencia Artificial & Homelab

Manual de Arquitectura de IA Local en Casa, Modelos LLM y Reconocimiento de Voz

Dimensionamiento de hardware x86, Ollama, Whisper, OCuLink eGPU y Home Assistant Voice.

OBJETO Y ALCANCE TÉCNICO
Especificación técnica para la implementación de modelos neuronales privados en local, garantizando 0 fuga de datos a la nube, latencia instantánea y control por voz sin cuotas.

1. Matriz de Requisitos de Hardware y VRAM para Modelos LLM

Modelo / Parámetros	RAM / VRAM Mínima	Hardware Recomendado	Tokens/s
Qwen 2.5 7B / 8B	16 GB DDR5 5600	Beelink SER8 (Ryzen 7)	~35 t/s
Qwen 2.5 14B / DeepSeek	32 GB DDR5 Dual	GEEKOM A8 (Ryzen 9)	~22 t/s
LLMs 32B / 70B + Visión	16GB VRAM Dedicada	GMKtec K8 + eGPU OCuLink	~40 t/s

2. Checklist de Despliegue y Aislamiento de Red 100% Local

- [ ] Instalar contenedor Docker de Ollama en red aislada (VLAN IoT) sin acceso saliente a Internet.
- [ ] Asignar memoria GPU compartida mínima de 8GB en BIOS (UMA Frame Buffer) para Radeon 780M / Intel Arc.
- [ ] Configurar Faster-Whisper con aceleración NPU para reconocimiento de voz en menos de 250ms.
- [ ] Enlazar con Home Assistant Voice Assist definiendo herramientas domóticas y restricciones de seguridad.