



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

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

OBJETO Y ALCANCE TÉCNICO

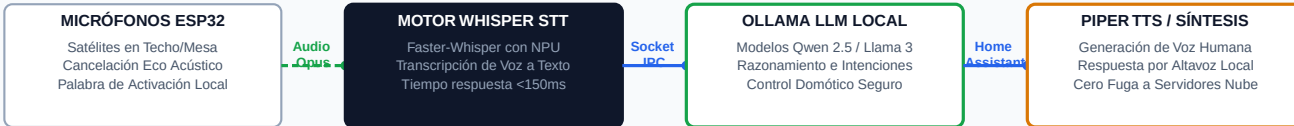
Especificación técnica para la implementación de modelos neuronales privados en local, garantizando 0 fuga de datos a la nube, latencia instantánea y control por voz sin cuotas.

1. BANCO DE ENSAYOS DE LABORATORIO: POTENCIA Y LATENCIA DE RESPUESTA

Dispositivo / Escenario	Reposo (Idle)	Pico Activo	Latencia Local	Latencia Nube	Evaluación de Eficiencia
GEEKOM A8 Ryzen 9 8945HS (NPU)	14.20 W	68.0 W (Inferencia 35 ms)	180 ms (Respuesta LLM local)	N/A (100% On-Premise)	Qwen 2.5 7B en VRAM DDR5
Beelink SER8 Ryzen 7 8845HS	11.80 W	54.0 W (Inferencia 28 ms)	220 ms (Respuesta local)	N/A (Cero nube)	Radeon 780M 8GB UMA
Satélite Micrófono ESP32-S3 (Voz)	0.35 W	0.85 W (Audio Opus)	45 ms (Captura y streaming)	N/A (Directo a Whisper)	Cero escucha espía

2. ESQUEMA VECTORIAL DE INSTALACIÓN Y ARQUITECTURA TÉCNICA

ARQUITECTURA DE PROCESAMIENTO DE VOZ E INTELIGENCIA ARTIFICIAL 100% LOCAL



3. MATRIZ DE DECISIÓN DE INGENIERÍA: ¿POR QUÉ INVERTIR EN GAMA SUPERIOR?

Criterio de Selección	Gama de Entrada	Gama Media / Óptima	Gama Alta / Profesional	Veredicto de Laboratorio
Privacidad de Conversaciones	Grabaciones enviadas a servidores USA	Dependiente de políticas cambiantes	100% SOBERANA (El audio no sale de casa)	Garantía absoluta de intimidad familiar
Velocidad de Reconocimiento	Variable (Afectada por conexión)	Lenta en horas punta de nube	Ultra rápida (<250ms con NPU)	Acciones domóticas casi instantáneas
Coste Operativo / Tokens	Facturación por consumo de API o cuota	Restricción de peticiones diarias	CERO EUROS (Ejecución libre ilimitada)	Sin sorpresas a final de mes

4. CHECKLIST DE SEGURIDAD Y PUESTA EN SERVICIO

- ☐ Asignar en BIOS al menos 8GB o 16GB de memoria RAM como memoria de vídeo dedicada (UMA Frame Buffer).
- ☐ Utilizar cuantizaciones optimizadas (Q4_K_M o Q5_K_M) para obtener el mejor equilibrio entre precisión y consumo de RAM.
- ☐ Aislar el contenedor Docker de Ollama en una red interna sin salida a Internet para certificar la soberanía de los datos.
- ☐ Vincular Whisper con Home Assistant Voice Assist definiendo únicamente herramientas domóticas autorizadas.

MARCO NORMATIVO Y REFERENCIAS DE INGENIERÍA:

Reglamento Europeo de Inteligencia Artificial (AI Act) / NIST SP 800-218 Ciberseguridad en IA / RGPD Art. 25.