

A voice AI platform in production. One engineer.

An AI assistant that answers phone calls, WhatsApp and web chat for pharmacies in Italy — each site speaking as itself, from its own catalogue, services, hours and staff. **Sixteen live deployments across eight regions**, one instance and one database per customer. Architected, built and operated end to end: voice pipeline, data layer, admin application, telephony and deployment.

SITES LIVE	TENANCY	CHANNELS	TEAM
16 eight regions	1 : 1 instance and database per customer	3 voice · WhatsApp · web chat	1 sole engineer, end to end

FOUR DECISIONS THAT MATTERED

01 One deployment per customer, not multi-tenancy

No tenant identifier on any table, no row-level scoping to get wrong — isolation by construction. The cost is fleet drift, paid in discipline: one canonical repository, every fix landing there first, backports tracked as a numbered queue.

02 Deleting the free-text prompt field

It was the largest source of behavioural drift: no bug reproduced, because no two sites said the same thing. Identity is now assembled deterministically from structured settings. A whole class of bug disappeared rather than being fixed.

03 Making the system refuse

The worst failure isn't silence — it's a fluent, specific, wrong answer. Confidence has to be earned from the data; below an explicit threshold, the assistant declines and hands the caller to a human.

04 Latency as architecture, not model choice

The model's share of the budget is fixed by the vendor, so everything else was won back in topology: same-region placement and the fewest possible hops between carrier and model.

The model was never the hard part. The four runtimes around it were — and keeping sixteen production deployments up while one person maintains them.

STACK	Bun / TypeScript · Laravel 12 + Filament 5 · Next.js · MySQL · Postgres + pgvector · Redis · Python · SIP/PSTN telephony · real-time speech models, streaming TTS, vision
-------	---

Richard Trujillo · Desemax. Senior backend and AI-integration engineer. Eleven years of backend engineering, independent since 2015. Based in Milan, working across Europe.

richard@desemax.it · richardtrujillo.desemax.it

Full case study, with architecture and component breakdown: richardtrujillo.desemax.it/case-studies