KUMMALABS · AI AGENT ENGINEERING · CASE STUDY

KOTA — turning restaurant phone calls into confirmed, structured orders (English + 中文)

A real-time voice agent that answers the call, takes the order against the restaurant's own menu data, records it, and confirms it to the caller.

What it does

The caller presses 1 for English or 2 for Chinese. The agent answers menu questions from the menu database, takes the order (standard items, combos, required options, quantity changes, removals, cart review), texts a menu link on request, places the order as a structured record, and texts the caller a confirmation. Pressing 0 at any time transfers the call to a staff phone.

How the work flows

Understand the input	Streaming speech-to-text on Twilio media streams — Deepgram (English), ByteDance (Mandarin)
Retrieve information	Menu, prices, options, and combos come from PostgreSQL; exact answers must come from tool calls
Apply business rules	Required options, combo structure, and cart operations enforced in code, not by prompt
Call the right tools	Typed tool calls: menu search, option selection, cart edit/remove, order summary, menu-link SMS, place order
Verify the action	The agent only says "I texted your order" if the messaging API reports success
Handle failure	No exact menu match → real menu suggestions instead of a guess; messaging failures are caught and logged without losing the order
Audit trail	Every call, utterance, and placed order is stored with the call ID
Escalate	Press 0 → the live call is redirected to a person

Engineering decisions

- **The model is never the source of truth.** Prices, items, and order state are reached only through deterministic tools — the main defense against an agent inventing a dish or a price.
- **Validated menu onboarding.** Menu photos/text → LLM extraction → grounding checks that remove pairings, recommendations, or aliases referencing items not on the menu → structure-preserving refinement → database ingest.
- **Multi-restaurant by design.** Prompt, persona, and menu are assembled per restaurant from the database.
- **Tested where it breaks.** 111 automated tests: tool contract, language routing, order options, menu grounding and search, messaging.

Stack

- Twilio voice + media streams, SMS, WhatsApp
- Deepgram streaming agent (English); ByteDance STT/TTS + OpenAI tool calling (Mandarin)
- Python, FastAPI WebSocket runtime, one handler per call
- PostgreSQL: restaurants, menus, calls, utterances, orders
- Docker Compose; Kubernetes manifests for Google Cloud

Scope boundaries

Not part of the current system: POS integration, kitchen printing, restaurant-side order dashboard, staff notification, payments, calendar booking, CRM sync. Each is a bounded integration on top of the structured order record.

> **Status, stated plainly:** a working, tested prototype built against real restaurant menus (combos, required options, bilingual item names). No restaurant has used it in production, it is not currently deployed, and we make no claims about call volume, accuracy, latency, or uptime.

Built by Kumma (Yang Wu) · KummaLabs — we build, connect, and improve AI agents that perform real work inside business systems · info@kummalabs.com