

Minicurso: Orquesta tu IA para Hostelería

Objetivo: que entiendas y montes un sistema de IA moderno sin mareos: agentes que colaboran, RAG para usar tus documentos, y evaluación continua para que nada se desmadre.

Para quién: dueños/gestores de restaurantes, hoteles, caterings, dark kitchens y pymes. Resultado: un flujo listo para producción “pedido-reserva-soporte” con coste controlado y trazabilidad.

Vocabulario express

- Agente: “personaje” que sabe hacer algo (reservas, traducir, mirar stock).
- Orquestación: cómo se reparten el trabajo varios agentes sin pisarse.
- RAG: el modelo consulta tus documentos (cartas, alérgenos, fichas) antes de responder.
- Evaluación continua: tests y métricas que vigilan calidad y coste 24/7.
- Function/Tool calling: el modelo llama funciones de tu sistema (TPV, CRM, ERP).

De chatbot a orquesta (la idea clave)

- Antes: un solo bot que habla bonito.
- Ahora: varios agentes especializados, con reglas claras para pasar el turno.
- Beneficio: respuestas más precisas, menos coste, menos lío.

Arquitectura mínima (que sí funciona)

- Cliente → Gateway (auth) → Orquestador → Agentes: Reservas, Carta/Alérgenos (RAG), Traducción, Resumen → CRM, Compliance (PII off).
- Regla de oro: cada agente hace una cosa bien. El orquestador decide quién actúa.

RAG para hostelería en 5 pasos

- Recolecta documentos: carta, alérgenos, escandallos, horarios, políticas.
- Limpia y trocea (chunks 300–800 palabras) con metadatos (fecha, versión, idioma).
- Vectoriza (embeddings) y sube a tu VectorDB (pgvector, Weaviate...).
- Recupera 3–8 trozos por consulta con filtros por versión/idioma.
- Genera respuesta citando fuente interna (Carta v3, 2025-07-15). Tip: RAG en streaming para menos latencia.

Evaluación & métricas (ahorra disgustos)

- Pruebas diarias: 20–50 preguntas fijas (alérgenos, reservas complejas, horarios).
- Reglas automáticas: si confianza < 0,3 → humano; si >100 tokens → exigir fuentes.
- KPIs: CSAT, FCR, coste/conversación, latencia p95, % escalado, % con fuente.

Privacidad y legal

- PII off: ofusca teléfonos y emails antes de guardar.
- Retención: define tiempos claros (p. ej., 90 días).
- Transparencia: muestra cuándo responde IA y cómo pedir revisión humana.

Caso práctico: Restaurante VivaPaella

- Objetivo: reservas y dudas de alérgenos 24/7 en ES/EN.
- Flujo: Cliente → Routing → (Reservas | Carta) → Resumen/CRM → Compliance.
- Métricas: 95% FCR (reservas simples), p95 < 15s, < 0,25€ por conversación.

Montaje en 1 tarde (checklist)

- Define 6–8 funciones (reservas, horarios, alérgenos, quejas, delivery).
- Sube RAG versionado (Carta_ES_v3_2025-07-15, Alérgenos_v2).
- Dibuja el grafo: Routing → (Reservas | Carta | Traducción) → Resumen.
- Activa guardrails (cita o no respondas, límite de tokens, PII off).
- Paneles: coste/latencia por agente, tasa “no sé”.
- Pruebas: 30 preguntas típicas (celiacos, grupos, cambios).

Prompts listos

- Orquestador: clasifica intención {reserva, carta, queja, delivery, otro} y enruta. Si ambiguo, UNA aclaración breve.
- Reservas (tools): usa funciones disponibles; si no hay mesa exacta, ofrece 2 alternativas; confirma nombre y móvil antes de crear.
- Carta/Alérgenos (RAG): responde solo con fragmentos versionados; si falta info → “no sé”.
- Resumen: 2 líneas con (intención, resultado), idioma y etiquetas.

Objetivos semanales

- 0 alucinaciones (si no hay fuente, no respondas).
- < 0,25 € coste/conversación; p95 < 15s; >90% FCR.

- 50 tests verdes; carta/alérgenos actualizados con fecha y versión.

TL;DR

Pasa de un chatbot simpático a una orquesta de agentes: orquestador claro, RAG con documentos versionados, evaluación continua y guardrails. Con esto, reservas y alérgenos quedan en automático y tú controlas calidad, coste y riesgo.