Propuesta de Caso Organizacional – Atención al Cliente en Hospital Público mediante Agentes de IA

Estudiantes: Allan Nuñez

Profesor/a: Giocrisrai Godoy Bonillo

Institución: Duoc UC

Asignatura: Ingeniería de soluciones con inteligencia artificial

Fecha: 22-09-2025

**1. Caso elegido.**

El caso el cual se eligió fue el de un hospital público el cual es El Hospital Clínico San Borja Arriarán es un centro de salud público ubicado en Santiago de Chile. “Atiende a una población estimada de más de 500.000 personas en el área metropolitana, ofreciendo servicios ambulatorios, hospitalización, urgencias y especialidades médicas.” Como institución pública, enfrenta una alta demanda y recursos limitados, lo que hace necesario innovar en sus procesos de atención y gestión administrativa.

**2. Problemática del caso.**

El hospital recibe diariamente cientos de solicitudes de pacientes relacionadas con información administrativa tal como estado de horas médicas, disponibilidad de fármacos, reclamos, solicitudes de reprogramación. Estas llegan a través de múltiples medios como por teléfono, correo electrónico y de manera presencial.

El proceso de atención es principalmente manual, lo que genera:

* Altos tiempos de espera para consultas frecuentes.
* Duplicación o pérdida de solicitudes por falta de trazabilidad.
* Sobrecarga de funcionarios administrativos por clasificación y derivación manual.

**3. Objetivos esperados.**

- se busca reducir un cierto porcentaje por ejemplo un 60% del tiempo de respuesta a consultas frecuentes de pacientes.

- Automatizar la clasificación y derivación de las solicitudes hacia las áreas correspondientes (farmacia, horas médicas, hospitalización, urgencias).

- Mejorar la trazabilidad de reclamos y consultas, garantizando que más del 95% queden registrados en un sistema digital.

- Optimizar el tiempo del personal administrativo, liberando un 20% de su carga para atención presencial crítica.

**4. Datos disponibles o que se pueden obtener**

* Registros históricos de consultas y reclamos por ejemplo los correos electrónicos, formularios, registros telefónicos.
* Categorías de atención establecidas en la gestión hospitalaria.
* Protocolos internos de atención al paciente y normativa ministerial.
* Datos simulados para entrenamiento inicial.

**5. Restricciones o requerimientos particulares**

* Restricción en el uso de datos sensibles de salud.
* Limitaciones de infraestructura tecnológica en un hospital público, que requieren soluciones escalables y de bajo costo.

**6. Motivación para el uso de agentes de IA, LLMs y RAG**

El problema identificado en la atención al cliente del hospital se caracteriza por la gran cantidad de solicitudes en lenguaje natural que llegan de manera desordenada y por múltiples canales. Estas solicitudes requieren ser entendidas, clasificadas y respondidas con rapidez y precisión. Una solución tradicional basada únicamente en formularios o sistemas de tickets no resuelve completamente la complejidad del escenario, ya que:

No interpreta el lenguaje natural de los pacientes que muchas veces es un lenguaje informal, con errores ortográficos o expresiones coloquiales.

No integra de manera automática la normativa vigente del Ministerio de Salud ni la documentación interna del hospital.

No ofrece respuestas inmediatas y personalizadas a dudas frecuentes, lo que genera tiempos de espera excesivos.

Ante esto, una arquitectura basada en agentes de IA resulta especialmente adecuada:

**Agentes de IA:**

Actúan como “coordinadores inteligentes” que deciden qué hacer con cada solicitud: responder directamente, consultar una base de datos, activar el pipeline RAG o escalar el caso a un funcionario humano.

Permiten automatizar flujos de trabajo que antes requerían de múltiples funcionarios administrativos.

**LLMs:**

Entienden la consulta del paciente en lenguaje natural, incluso si está escrita con errores o de manera poco estructurada.

Generan respuestas claras, comprensibles y adaptadas al nivel de comprensión del usuario.

Pueden manejar contextos de conversación, lo que simula un “diálogo humano” y aumenta la satisfacción del paciente.

**Pipeline RAG:**

Permite que el LLM no dependa solo de su entrenamiento, sino que acceda a información actualizada y específica del hospital.

Combina fuentes internas como los protocolos de atención, manuales, categorías de reclamos con fuentes externas.

Garantiza que las respuestas sean precisas, contextuales y verificables, reduciendo el riesgo de respuestas incorrectas o inventadas.

En conjunto, esta arquitectura ofrece una solución que no solo mejora la atención al paciente, sino que también:

Reduce costos operativos al liberar tiempo de los funcionarios.

Aumenta la satisfacción de los usuarios, al obtener respuestas rápidas y confiables.

Fortalece la transparencia y trazabilidad, ya que cada interacción queda registrada en el sistema digital del hospital.

Escala fácilmente a nuevos servicios o áreas del hospital, gracias a la modularidad de los agentes y el pipeline RAG.

### **7. Implementación Técnica: Mecanismos de Memoria y Recuperación de Contexto**

**Memoria de Contenido:** Este mecanismo permite mantener la coherencia durante una misma interacción, recordando el historial reciente de la conversación entre el paciente y el agente. Se implementa mediante un buffer de conversación que almacena los últimos intercambios, evitando repeticiones o pérdidas de información.  
 Gracias a esto, el agente puede continuar flujos prolongados, como el registro de un reclamo, comprendiendo el contexto sin necesidad de que el usuario repita la información.

**Resultado esperado:** mayor continuidad y coherencia en las interacciones.

**Recuperación de Contexto Semántico:** La memoria de largo plazo se implementa con un Pipeline RAG, que permite al agente acceder a una base de conocimiento con información actualizada del hospital (protocolos, normativas y manuales).  
 Cuando un paciente realiza una consulta, el sistema busca los documentos más relevantes y los utiliza para generar una respuesta precisa y verificada.

**Resultado esperado:** coherencia y precisión en las respuestas, basadas en información real del hospital y no solo en el entrenamiento previo del modelo.

### **8. Esquemas de Planificación y Toma de Decisiones**

**1. Planificación y Secuenciación de Tareas:** El agente utiliza un esquema de planificación que determina si una solicitud debe resolverse automáticamente o derivarse a un funcionario.  
 Las tareas simples, como preguntas frecuentes, se responden de forma automática mediante el uso del pipeline RAG, mientras que las solicitudes más complejas o que requieren intervención humana se escalan al área correspondiente.  
 Ejemplos:

* *Preguntas frecuentes:* se procesan directamente, optimizando tiempos de respuesta.
* *Reprogramaciones de citas:* se inician flujos administrativos automatizados para posterior revisión.
* *Reclamos o casos complejos:* se derivan a personal humano sólo si el sistema no puede resolverlos, garantizando eficiencia y precisión.

**2. Toma de Decisiones Adaptativa:** El agente ajusta su comportamiento según las condiciones del entorno o el éxito del sistema RAG:

* **Caso 1:** Si el RAG encuentra información (por ejemplo, horarios de visita), el agente responde automáticamente.
* **Caso 2:** Si el RAG no aplica (por ejemplo, en un reclamo), el agente deriva el caso a un funcionario.
* **Caso 3:** En consultas de urgencia, el RAG entrega respuestas basadas en protocolos médicos, asegurando una orientación inmediata y segura.

**Resultado esperado:** Un sistema capaz de tomar decisiones autónomas, reducir la carga administrativa y mejorar la calidad y rapidez de atención al paciente.

### 

### 

### 

### **9. Orquestación de Componentes y Justificación Arquitectónica**

**1. Diagrama de Orquestación de Componentes:**  
 El proceso inicia cuando el paciente envía una consulta al Agente Coordinador, el cual clasifica la intención del usuario y decide el flujo de acción.

La Memoria de Corto Plazo mantiene la coherencia de la conversación, mientras que el Pipeline RAG recupera información relevante desde la base de conocimiento hospitalaria.  
 Según el tipo de consulta, el agente puede:

* Generar una respuesta automática si cuenta con contexto suficiente.
* Derivar la solicitud a personal humano si el caso requiere acción directa.  
   Finalmente, la Herramienta de Escritura formula la respuesta y actualiza el registro en la memoria, asegurando trazabilidad y continuidad.

**2. Justificación de los Componentes:** La arquitectura modular fue diseñada para resolver los principales problemas del Hospital San Borja Arriarán: tiempos de espera, trazabilidad y sobrecarga administrativa.

* **Agente Coordinador:** Gestiona la planificación y toma de decisiones, reduciendo la carga de los funcionarios.
* **Pipeline RAG:** Asegura respuestas precisas y verificables mediante el acceso a información actualizada.
* **Memoria de Contenido:** Mantiene coherencia en interacciones prolongadas, mejorando la experiencia del paciente.
* **LLM (Modelo de Lenguaje):** Interpreta lenguaje natural y redacta respuestas comprensibles y contextualizadas.

**10. Flujos de trabajo:**

#### **Flujo de Trabajo: Consulta Frecuente**

| **Paso** | **Módulo del Agente** | **Acción / Decisión** | **Objetivo** |
| --- | --- | --- | --- |
| **Entrada** | N/A | El paciente pregunta: *"¿Cuál es el horario de visita?"* | Capturar la solicitud. |
| **Clasificación** | clasificacion() | El Agente clasifica la intención como pregunta\_frecuente. | Determinar el camino de alta prioridad. |
| **Consulta** | contexto(RAG) | Se busca la información de "horarios de visita" en la Base de Conocimiento. | **Condición:** RAG exitoso. |
| **Escritura y Decisión** | generador\_respuesta() | Utiliza el contexto recuperado del RAG para generar la respuesta. | **Decisión:** Responder automáticamente. |
| **Salida y Memoria** | update\_memory() | Responde y registra la interacción en el buffer de memoria. | Optimización: Evita la derivación humana. |

#### **Flujo de Trabajo : Gestión de Reclamos**

| **Paso** | **Módulo del Agente** | **Acción / Decisión** | **Objetivo** |
| --- | --- | --- | --- |
| **Entrada** | N/A | El paciente dice: *"Tengo un reclamo por un problema en la atención."* | Capturar la solicitud de acción. |
| **Clasificación** | clasificacion() | El Agente clasifica la intención como reclamo. | Determinar el camino de derivación. |
| **Consulta** | contexto(RAG) | Se ignora la búsqueda RAG o se usa para obtener el *protocolo de reclamos* (no una respuesta directa). | La acción es más importante que la información. |
| **Decisión de Planificación** | process\_query() | **Decisión:** Derivación inmediata a funcionario, independientemente del contexto. | Garantizar trazabilidad y registro formal. |
| **Escritura y Memoria** | generador\_respuesta() | Genera una respuesta de confirmación de escalamiento y registra el inicio del proceso. | Informar al paciente y registrar la interacción. |

**11. Referencias y Frameworks usados:**

**Lenguajes de Programación y Librerías de Orquestación:**

* **LangChain.** (s.f.). LangChain: Building applications with LLMs through composability. Recuperado de<https://www.langchain.com/>
* \_Diario Financiero. (2024). Hospitales públicos enfrentan aumento en reclamos por tiempos de espera. <https://www.df.cl>

Ministerio de Salud de Chile. (2023). Normativa de atención hospitalaria y derechos de los pacientes. MINSAL.