



Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede

Trabajo fin de estudio presentado por:	Mª Gemma Pascual Santos José Antonio Pérez Criado
Tipo de trabajo:	TFM-Inteligencia Artificial
Director/a:	Deidy Gutiérrez Niño
Fecha:	26 marzo de 2026

Índice

1	Resumen	3
2	Abstract.....	3
3	Introducción	5
3.1	Objetivo General	5
3.2	Objetivos Específicos.....	5
3.3	Metodología	6
3.3.1	Identificación y Análisis del Problema	6
3.3.2	Planificación de la Solución.....	6
4	Identificación y descripción un problema al interior de la organización	8
4.1	Descripción de la empresa a intervenir.....	8
4.1.1	Misión	8
4.1.2	Visión	8
4.1.3	Mapa de procesos	8
4.1.4	Descripción del proceso a intervenir	9
4.1.5	Descripción del problema que presenta el proceso	10
5	Establecimiento de planes de acción para la solución del problema	12
5.1	Establecimiento de objetivos.....	12
5.2	Descripción de la solución	13
5.3	Secuencia de actividades para ejecución de la solución	15
5.4	Descripción de recursos necesarios para la ejecución de la solución	17
6	Definición del alcance de la solución, partes interesadas, cronograma y presupuestos. 20	
6.1	Alcance final de la solución	20
6.2	Identificación de responsables de las actividades para ejecución.....	20
6.3	Cronograma de ejecución.....	21
6.4	Presupuesto	22
7	Referencias bibliográficas	31

1 Resumen

El presente trabajo aborda una problemática identificada en la gestión de información y atención de consultas en una clínica privada de servicios médicos especializados. Actualmente, una gran parte de las consultas de pacientes y del personal administrativo se gestionan mediante atención telefónica o mediante la búsqueda manual de información en diferentes documentos internos, lo que genera tiempos de respuesta elevados y una carga operativa considerable.

Con el objetivo de mejorar la eficiencia en el acceso a la información, se propone el diseño e implementación de un asistente conversacional basado en inteligencia artificial que utilice una arquitectura de generación aumentada por recuperación (Retrieval-Augmented Generation, RAG). Este sistema permitirá consultar de forma automática información institucional, protocolos y servicios médicos a través de lenguaje natural, facilitando el acceso rápido y organizado a la información disponible.

La propuesta incluye la creación de un sistema de recuperación de información que permita consultar la documentación institucional de la clínica. Para ello se utilizarán modelos de lenguaje capaces de generar respuestas basadas en dicha información. Además, se incorporarán mecanismos básicos de control que ayuden a asegurar que las respuestas del asistente estén fundamentadas en los documentos disponibles.

Con la implementación de la propuesta, se pretende mejorar los tiempos de respuesta en la atención de consultas, reducir la carga operativa del personal administrativo y facilitar el acceso a la información tanto para el personal de la clínica como para los pacientes.

Además, la propuesta demuestra que este tipo de soluciones puede aplicarse de forma realista en entornos sanitarios, aportando mejoras en la gestión de la información y en la atención al paciente sin requerir grandes cambios en la infraestructura existente.

Palabras clave: arquitectura RAG, asistentes virtuales, inteligencia artificial, procesamiento de lenguaje natural, sistemas conversacionales.

2 Abstract

This study addresses a problem identified in the management of information and the handling of inquiries in a private healthcare clinic that provides specialized medical services. Currently, a large proportion of patient and administrative queries are managed through telephone communication or through manual searches across different internal documents, which leads to long response times and a significant operational workload.

In order to improve the efficiency of information access, this project proposes the design and implementation of a conversational assistant based on artificial intelligence using a Retrieval-Augmented Generation (RAG) architecture. This system will allow users to automatically consult institutional information, medical protocols, and service details through natural language interactions, facilitating faster and more organized access to available information.

The proposal includes the creation of an information retrieval system that allows users to consult the clinic's institutional documentation. Language models will be used to generate responses based on this information. In addition, basic control mechanisms will be incorporated to ensure that the assistant's responses are supported by the available documentation.

With the implementation of this proposal, the aim is to improve response times for handling inquiries, reduce the administrative workload, and facilitate access to information for both clinic staff and patients.

In addition, this proposal shows that this type of solution can be realistically applied in healthcare environments, contributing to improved information management and patient care without requiring significant changes to the existing infrastructure.

Keywords: artificial intelligence, conversational systems, natural language processing, RAG architecture, virtual assistants.

3 Introducción

Este proyecto de aplicación en investigación-acción se centró en la identificación y resolución de las ineficiencias en la gestión de información interna y en el proceso de asignación de citas médicas en la Clínica Salud Integral, permitiendo integrar teoría y práctica en un entorno real. Mediante un enfoque de investigación-acción, se desarrollaron habilidades de diagnóstico, análisis y planificación orientadas a mejorar la eficiencia y efectividad de un proceso crítico en la organización. La estructura del proyecto incluyó una revisión exhaustiva del contexto organizacional y la implementación de una solución factible que responde a las necesidades de la organización y optimiza su desempeño.

En los últimos años, los sistemas conversacionales basados en inteligencia artificial han empezado a utilizarse cada vez más en entornos organizacionales, especialmente en el ámbito sanitario, donde la gestión adecuada de la información es fundamental. Diversas investigaciones recientes señalan que la combinación de modelos de lenguaje con sistemas de recuperación de información permite generar respuestas más precisas y mejor fundamentadas en datos internos (Lewis et al., 2020; Gao et al., 2023; Zhang et al., 2022). Además, se ha observado que la incorporación de asistentes virtuales en el sector salud puede ayudar a reducir la carga administrativa y mejorar la experiencia del paciente al facilitar la resolución de consultas frecuentes de manera más ágil.

En este contexto, el uso de arquitecturas basadas en Retrieval-Augmented Generation (RAG) se presenta como una alternativa prometedora para mejorar la recuperación de información y garantizar que las respuestas generadas estén fundamentadas en documentación institucional.

3.1 Objetivo General

El objetivo principal del proyecto fue desarrollar una solución operativa al problema de las ineficiencias en la gestión de información interna y en el proceso de asignación de citas médicas en la Clínica Salud Integral. La intervención se orientó no solo a corregir el problema identificado, sino también a proporcionar una propuesta sostenible que pudiera integrarse a la estructura de la organización, elevando los niveles de eficiencia y alcanzando los estándares de desempeño requeridos en el proceso intervenido.

3.2 Objetivos Específicos

Para cumplir con el objetivo general, se establecieron tres objetivos específicos:

1. Realizar una identificación detallada del problema en su contexto organizacional, determinando causas raíz y evaluando su impacto en el desempeño general del proceso.
2. Diseñar un plan de acción estructurado con objetivos claros, actividades específicas y recursos definidos para la ejecución de la solución.

3. Establecer el alcance de la intervención, definiendo responsables, cronograma y presupuesto, para asegurar que la implementación sea organizada y efectiva.

3.3 Metodología

El proyecto se desarrolló mediante la metodología de investigación-acción, la cual facilitó la colaboración continua entre el equipo de estudiantes y docente asesor. Esta metodología permitió un enfoque sistemático para identificar problemas, formular hipótesis y aplicar soluciones en un contexto real, ajustando las estrategias conforme se avanzaba en el proyecto. La investigación-acción proporcionó un marco flexible y participativo para ejecutar el proyecto de manera estructurada, generando una intervención bien fundamentada y con alto potencial de impacto.

El desarrollo del proyecto se estructuró en tres etapas secuenciales, cada una de las cuales abordó un aspecto específico de la intervención. En la primera etapa, se identificó y analizó el problema organizacional, incluyendo una descripción detallada de la organización, su misión, visión y el proceso específico afectado. La segunda etapa se centró en la planificación de la solución, estableciendo objetivos y actividades concretas para abordar el problema. En la tercera fase, se definieron aspectos operativos como el cronograma, el presupuesto y la asignación de roles.

3.3.1 Identificación y Análisis del Problema

En la primera etapa, se realizó un análisis exhaustivo de la organización y del contexto en el cual se desarrolla el proceso afectado. Se documentó el problema identificando sus causas raíz, efectos y puntos críticos que afectaban el desempeño del proceso. Este diagnóstico inicial fue esencial para estructurar la solución y asegurar que la intervención respondiera de manera precisa a las necesidades de la organización.

Para la identificación del problema se utilizó un enfoque de análisis organizacional basado en revisión documental interna, observación del proceso de atención y entrevistas informales con el personal administrativo encargado de la gestión de citas. Este diagnóstico permitió identificar tiempos de respuesta elevados, duplicidad de información y dependencia excesiva de la atención telefónica para resolver consultas frecuentes. A partir de este análisis se determinaron los puntos críticos del proceso y se estableció la necesidad de una solución tecnológica que optimizara la gestión de la información y redujera la carga operativa del personal.

3.3.2 Planificación de la Solución

En la segunda etapa, se desarrolló un plan de acción detallado que incluyó la definición de objetivos específicos y una secuencia de actividades que permitiría mitigar o resolver el problema identificado. Esta fase incluyó el diseño de una estrategia de ejecución organizada y estructurada que facilitara el logro de los objetivos propuestos y maximizara la eficiencia del proceso intervenido. La planificación detallada de actividades y recursos fue clave para asegurar una implementación precisa y efectiva.

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede.

3.3.2.1 Definición de Alcance y Recursos

La tercera etapa se enfocó en delimitar el alcance de la intervención y los recursos necesarios. Se establecieron responsables para cada actividad, y se elaboró un cronograma de ejecución para controlar el avance de la implementación. Asimismo, se desarrolló un presupuesto que detalló los costos asociados a cada fase, permitiendo optimizar la asignación de recursos. Esta fase aseguró que todas las variables críticas estuvieran alineadas para una ejecución sin interrupciones.

4 Identificación y descripción un problema al interior de la organización

En el contexto de la Clínica Salud Integral, institución privada dedicada a la prestación de servicios médicos especializados en distintas sedes, se ha identificado una problemática relacionada con la gestión de la información y la asignación de citas médicas. El proceso actual depende en gran medida de la atención telefónica y de la consulta manual de información interna por parte del personal administrativo, lo que genera tiempos de respuesta elevados y una carga operativa significativa.

La información relevante sobre servicios, horarios, disponibilidad de especialistas y políticas internas se encuentra distribuida en distintos documentos y sistemas, lo que dificulta el acceso rápido y uniforme a los datos. Como consecuencia, se presentan inconsistencias en las respuestas ofrecidas a los pacientes y una dependencia constante del personal para resolver consultas que podrían ser automatizadas.

Esta situación afecta directamente la eficiencia del proceso de atención, incrementa el tiempo de espera y limita la capacidad del equipo administrativo para enfocarse en tareas de mayor complejidad. Por ello, se considera necesario intervenir este proceso con una solución tecnológica que permita optimizar la gestión de la información y mejorar la experiencia del paciente.

4.1 Descripción de la empresa a intervenir

4.1.1 Misión

La misión de la Clínica Salud Integral es brindar servicios médicos especializados con calidad, eficiencia y trato humano, garantizando una atención oportuna y centrada en las necesidades del paciente.

4.1.2 Visión

La visión de la Clínica Salud Integral es consolidarse como una institución referente en servicios de salud a nivel local, destacándose por la innovación en sus procesos, la mejora continua y la implementación de soluciones tecnológicas que optimicen la experiencia del paciente.

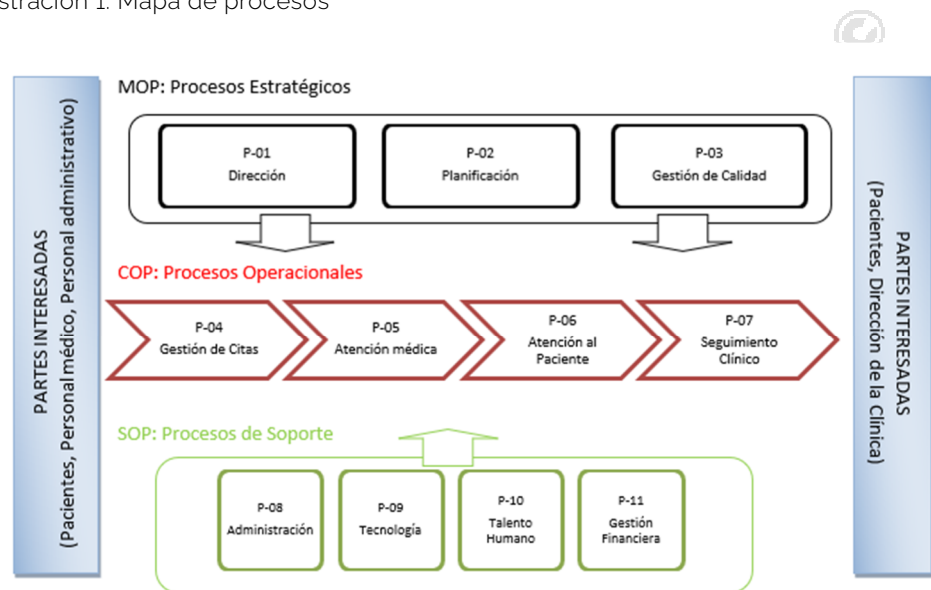
4.1.3 Mapa de procesos

La Clínica Salud Integral organiza sus actividades a través de procesos estratégicos, misionales y de apoyo. Los procesos estratégicos están relacionados con la dirección y

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede.

planificación institucional; los procesos misionales comprenden la atención médica, la gestión de citas y la atención al paciente; y los procesos de apoyo incluyen áreas como administración, tecnología y gestión del talento humano. Dentro de este esquema, el proceso de gestión de información y asignación de citas forma parte de los procesos misionales, ya que impacta directamente en la experiencia del paciente y en la eficiencia operativa de la organización.

Ilustración 1. Mapa de procesos



Fuente: Elaboración propia

4.1.4 Descripción del proceso a intervenir

El proceso a intervenir corresponde a la gestión de información y asignación de citas médicas en la Clínica Salud Integral. Actualmente, este proceso se desarrolla principalmente a través de atención telefónica y presencial, donde el personal administrativo consulta la disponibilidad de especialistas, horarios y servicios en los sistemas internos de la clínica para programar las citas solicitadas por los pacientes.

Cuando un paciente requiere información sobre servicios, requisitos o disponibilidad, debe comunicarse directamente con la clínica, lo que implica que cada consulta sea atendida de manera manual. El personal debe verificar la información en diferentes documentos o plataformas internas antes de ofrecer una respuesta, lo que incrementa el tiempo de atención y genera dependencia constante del recurso humano.

Este proceso, aunque funcional, presenta limitaciones relacionadas con la eficiencia operativa, la estandarización de la información y la capacidad de respuesta ante un volumen elevado de consultas, especialmente en horarios de alta demanda.

4.1.5 Descripción del problema que presenta el proceso

El principal problema identificado en el proceso de gestión de información y asignación de citas médicas es la dependencia total del personal administrativo para resolver consultas frecuentes y programar citas. Esto genera una sobrecarga operativa, especialmente en horarios de alta demanda, donde el volumen de llamadas y solicitudes supera la capacidad de respuesta inmediata.

Además, la información relevante para atender a los pacientes se encuentra distribuida en distintos documentos y sistemas internos, lo que obliga al personal a consultar varias fuentes antes de proporcionar una respuesta. Esta situación incrementa los tiempos de atención y puede dar lugar a inconsistencias en la información suministrada.

Como consecuencia, se presentan demoras en la asignación de citas, saturación de los canales de comunicación y una experiencia de usuario poco ágil. El proceso actual no cuenta con mecanismos automatizados que permitan filtrar o resolver consultas básicas, lo que limita la eficiencia general del servicio y afecta indirectamente la percepción de calidad por parte de los pacientes.

Análisis DOFA del proceso de gestión de información y asignación de citas

Con el fin de identificar de forma estructurada los factores internos y externos que influyen en el proceso analizado, se elaboró una matriz DOFA. Esta herramienta permite analizar las fortalezas y debilidades del proceso dentro de la organización, así como las oportunidades y amenazas del entorno que pueden afectar su desempeño.

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede.

Tabla 1. Matriz DOFA del proceso de gestión de información y citas

FORTALEZAS	DEBILIDADES
• Experiencia del personal administrativo en la atención al paciente • Disponibilidad de información institucional sobre servicios médicos • Conocimiento profundo del proceso de atención por parte del personal	• Dependencia de la gestión manual de consultas y citas • Información distribuida en múltiples documentos y sistemas no integrados • Tiempos de respuesta elevados en horarios de alta demanda
OPORTUNIDADES	AMENAZAS
• Implementación de herramientas de IA para automatizar consultas frecuentes • Mejora de la eficiencia mediante sistemas conversacionales RAG • Optimización de la experiencia del paciente con disponibilidad 24/7	• Saturación de canales de comunicación en horarios pico • Riesgo de inconsistencias en la información proporcionada • Insatisfacción de los pacientes por demoras en la atención

Fuente: Elaboración propia

A partir del análisis realizado se observa que, aunque la organización cuenta con fortalezas como la experiencia del personal administrativo y la disponibilidad de información institucional, existen debilidades importantes relacionadas con la dependencia de procesos manuales y la dispersión de la información en diferentes documentos y sistemas.

Estas debilidades generan tiempos de respuesta elevados y una sobrecarga operativa en el personal encargado de la atención al paciente. Al mismo tiempo, el entorno presenta oportunidades relevantes asociadas al uso de tecnologías basadas en inteligencia artificial que permiten automatizar la gestión de consultas frecuentes y mejorar la eficiencia de los procesos de atención.

En este contexto, la implementación de un sistema conversacional basado en recuperación aumentada por generación (RAG) se presenta como una alternativa adecuada para aprovechar estas oportunidades, ya que permitiría centralizar la información institucional, automatizar consultas recurrentes y mejorar la gestión de citas médicas, contribuyendo así a optimizar la experiencia del paciente y la eficiencia operativa de la organización.

En términos prácticos, esto permitiría que tanto pacientes como personal de la clínica obtuvieran respuestas rápidas a sus consultas sin depender siempre de la atención telefónica.

5 Establecimiento de planes de acción para la solución del problema

A partir del diagnóstico realizado, se plantea un plan de acción integral orientado a reducir la carga administrativa, estandarizar la información ofrecida a pacientes y empleados, y asegurar respuestas coherentes, verificables y trazables. El plan se basa en el desarrollo de un asistente conversacional basado en RAG con dos perfiles diferenciados —uno para el personal de la clínica y otro para los pacientes—. Además, se contemplan mecanismos de evaluación y medidas de seguridad y privacidad acordes con la normativa del entorno sanitario. (RGPD, Ley Orgánica 3/2018).

5.1 Establecimiento de objetivos

Los objetivos se definen siguiendo criterios SMART (específicos, medibles, alcanzables, relevantes y acotados en el tiempo). Se prioriza la mejora del tiempo de respuesta a consultas frecuentes, la reducción de llamadas repetitivas y la consistencia de la información proporcionada. Asimismo, se incorporan métricas técnicas de calidad RAG para garantizar que la solución sea evaluable y replicable.

Objetivo General:

Diseñar e implementar un asistente conversacional basado en RAG para la Clínica Salud Integral que permita resolver consultas internas (personal) y externas (pacientes) con respuestas fundamentadas en documentación oficial, incluyendo citación de fuentes, mecanismos de abstención y un marco de evaluación de calidad, seguridad y rendimiento.

Objetivos Específicos:

- Construir y curar un repositorio documental centralizado (servicios, protocolos, políticas, preparación de pruebas y FAQs), definiendo un esquema de metadatos, versiones y control de acceso por perfil de usuario.
- Desarrollar el sistema RAG completo, incluyendo la organización de la información, su indexación y los mecanismos de recuperación que permitan generar respuestas fundamentadas en la documentación disponible.
- Diseñar y ejecutar un plan de evaluación cuantitativa con un banco de preguntas (gold set) que mida precisión de recuperación (Precision@k, MRR), fidelidad de la respuesta (groundedness), robustez frente a prompt-injection, latencia y satisfacción del usuario (CSAT).
- Implementar guardrails de seguridad y privacidad que impidan la divulgación de datos personales protegidos, detecten intentos de manipulación del sistema y activen el protocolo de abstención cuando la documentación disponible sea insuficiente para fundamentar una respuesta.

5.2 Descripción de la solución

Perfil Interno — Personal de la Clínica

El perfil interno está diseñado para el personal de recepción, administración y sanitarios (médicos, enfermería y auxiliares). Este modo requiere autenticación mediante credenciales institucionales y ofrece acceso segmentado a documentación operativa sensible, incluyendo protocolos clínicos, normativas internas, guías de procedimientos, calendarios de disponibilidad de especialistas y registros de políticas de la organización.

Las funcionalidades clave del perfil interno son:

- Consulta de protocolos clínicos y procedimientos operativos estándar (POEs) con citación exacta del documento fuente, número de versión y fecha de actualización.
- Búsqueda semántica en el repositorio documental institucional mediante lenguaje natural, sin necesidad de conocer la ubicación exacta del documento.
- Asistencia en la gestión de citas complejas, verificando disponibilidad de especialistas, requisitos de preparación previa y posibles conflictos de agenda.
- Soporte para la resolución de dudas frecuentes del personal nuevo (onboarding), reduciendo la dependencia de consultas a supervisores para preguntas procedimentales.
- Generación de respuestas con trazabilidad completa: cada respuesta indica el fragmento documental recuperado, la puntuación de relevancia y el nivel de confianza de la respuesta.
- Activación de modo de abstención cuando la pregunta trasciende el alcance del repositorio disponible, derivando la consulta al responsable competente.

Perfil Externo — Pacientes

El perfil externo está orientado a pacientes que acceden a través del sitio web corporativo, WhatsApp Business u otros canales digitales de la clínica. Este modo no requiere autenticación y opera exclusivamente sobre información pública y de orientación: catálogo de servicios, horarios de atención, especialidades disponibles, ubicación y contacto de sedes, información de preparación para pruebas diagnósticas y preguntas frecuentes de carácter general.

De esta manera, los pacientes podrían resolver muchas de sus dudas habituales de forma inmediata, reduciendo tiempos de espera y mejorando la experiencia general con el servicio de la clínica.

Las funcionalidades clave del perfil externo son:

- Respuesta inmediata a consultas frecuentes sobre servicios, horarios, precios orientativos y ubicaciones de sedes, disponible las 24 horas del día.

- Guía paso a paso para la solicitud y gestión de citas médicas (nuevas citas, modificaciones y cancelaciones) integrando, en fases avanzadas, conexión con el sistema de agendamiento de la clínica.
- Información de preparación previa para pruebas diagnósticas (analíticas, pruebas de imagen, endoscopias, etc.) con instrucciones claras y citación de la fuente documental oficial.
- Derivación proactiva a atención humana cuando la consulta involucra síntomas agudos, urgencias médicas o situaciones que requieran valoración clínica, con mensaje explícito sobre la limitación del sistema.
- Protección reforzada de datos: el asistente externo no solicita ni procesa datos personales identificativos; cualquier dato necesario para la gestión de citas es gestionado por los sistemas back-end de la clínica conforme al RGPD.
- Respuestas formuladas en lenguaje accesible, evitando tecnicismos médicos innecesarios y adaptando el nivel de detalle al perfil del usuario.

Arquitectura técnica del sistema RAG

La arquitectura del sistema se organiza en cinco capas funcionales interconectadas que garantizan la calidad, seguridad y mantenibilidad de la solución:

Tabla 2. Capas funcionales de la arquitectura RAG

Capa	Componentes y función
Ingesta documental	Procesamiento y normalización de documentos (PDF, DOCX, HTML). Extracción de metadatos: tipo de documento, versión, fecha, audiencia objetivo (interno/externo), especialidad médica asociada.
Chunking semántico	División de los documentos en fragmentos más pequeños que mantengan el contexto de la información, facilitando su recuperación durante el proceso de búsqueda.
Indexación híbrida	Doble indexación: índice vectorial denso (embeddings) para búsqueda semántica, e índice BM25 para búsqueda léxica exacta. Almacenamiento en base vectorial (FAISS o Chroma) con metafiltrado por perfil de acceso.
Recuperación y re-ranking	Recuperación de los fragmentos más relevantes combinando distintos criterios de búsqueda para asegurar que la respuesta esté basada en la información más pertinente. Re-ranking mediante Cross-Encoder para reordenar los fragmentos según relevancia real a la consulta específica.
Generación y guardrails	Generación de respuesta mediante LLM con prompt estructurado que incluye instrucciones de citación y abstención. Guardrails de entrada (detección de prompt-injection, consultas fuera de alcance) y de salida (verificación de groundedness, filtrado de datos personales).

Fuente: Elaboración propia.

5.3 Secuencia de actividades para ejecución de la solución

La ejecución se organiza en actividades iterativas con entregables técnicos y operativos claramente definidos para cada hito del proyecto.

Actividad 1: Levantamiento y normalización del conocimiento

Para llevar a cabo esta fase se definieron varias tareas orientadas a organizar y estructurar la información institucional disponible.

Esta actividad comprende la recopilación, clasificación y normalización de todos los documentos institucionales relevantes para el funcionamiento del asistente. Se trata del fundamento del sistema RAG: sin una base documental bien estructurada y curada, la calidad de las respuestas se verá directamente comprometida.

Las tareas específicas incluyen:

1. Inventario documental: identificación y catalogación de todos los documentos existentes (servicios ofrecidos, horarios por sede y especialidad, protocolos de preparación para pruebas, políticas de cancelación y reprogramación, preguntas frecuentes y normativas internas).
2. Definición del esquema de metadatos: para cada documento se registrarán los campos tipo_documento, versión, fecha_creación, fecha_actualización, audiencia (interno/externo), especialidad_médica, sede_aplicable y nivel_confidencialidad.
3. Implementación del control de versiones: se establecerá un flujo de actualización documental con aprobación del responsable de cada área, garantizando que el repositorio refleje siempre la información vigente.
4. Validación de contenidos: revisión del personal clínico y administrativo para confirmar la exactitud, vigencia y completitud de los documentos incorporados al repositorio.
5. Implementación de la base vectorial: configuración de la base de datos vectorial con los índices híbridos (denso + BM25) y los filtros de acceso por perfil.

Entregable: repositorio documental estructurado con metadatos completos, control de versiones activo y acceso segmentado por perfil operativo.

Actividad 2: Desarrollo del pipeline RAG y guardrails

En esta etapa se desarrollan las principales funcionalidades técnicas del sistema conversacional.

Esta actividad comprende la implementación técnica del núcleo del sistema: el pipeline RAG completo y los mecanismos de seguridad y control de calidad de las respuestas.

Las tareas específicas incluyen:

- 1- Implementación del pipeline de ingesta: desarrollo del módulo de carga, parsing y chunking semántico con solapamiento contextual, utilizando LangChain o LlamaIndex como framework de orquestación.
- 2- Configuración del mecanismo de recuperación de información que permita localizar los documentos más relevantes dentro del repositorio.
- 3- Diseño de las plantillas de prompts: desarrollo de prompts diferenciados para el perfil interno y externo, con instrucciones explícitas de citación de fuentes, nivel de detalle apropiado y protocolo de abstención ante información insuficiente.
- 4- Implementación de guardrails de entrada: filtros de detección de prompt-injection, clasificador de intención para identificar consultas fuera del alcance del sistema y validador de idioma.
- 5- Implementación de guardrails de salida: verificador de groundedness (porcentaje de la respuesta fundamentado en los fragmentos recuperados), detector de información personal identificable (PII) y filtro de contenido médico de riesgo.
- 6- Desarrollo de la interfaz de usuario: implementación del frontend web para ambos perfiles, con gestión de sesiones autenticadas (perfil interno) y acceso anónimo seguro (perfil externo).
- 7- Integración con canales externos (fase opcional): configuración del conector con WhatsApp Business API para el perfil externo.

Entregable: sistema RAG funcional con ambos perfiles operativos, guardrails activos y documentación técnica del pipeline.

Actividad 3: Evaluación, pilotaje y ajuste

Con el fin de validar el funcionamiento del sistema, se plantean las siguientes actividades de evaluación y prueba.

Esta actividad garantiza que el sistema cumple los estándares de calidad definidos antes de su despliegue definitivo en producción.

Las tareas específicas incluyen:

1. Construcción del gold set de evaluación: desarrollo de un banco de preguntas representativo (mínimo 100 preguntas por perfil) con respuestas de referencia

validadas por el personal de la clínica, cubriendo consultas frecuentes, casos límite y preguntas adversarias.

2. Evaluación automática de métricas RAG: evaluación del sistema mediante métricas de calidad que permitan analizar la precisión de las respuestas, la relevancia de la información recuperada y el nivel de satisfacción de los usuarios.
3. Pruebas de seguridad: ejecución de intentos de prompt-injection, consultas de diagnóstico médico y solicitudes de datos personales para verificar la robustez de los guardrails.
4. Pruebas de rendimiento: medición de latencia extremo a extremo (objetivo: p95 < 3 segundos) y pruebas de carga concurrente.
5. Pilotaje con usuarios reales: prueba con un grupo reducido de personal (perfil interno) y pacientes voluntarios (perfil externo) durante dos semanas, recopilando feedback mediante encuestas de satisfacción (CSAT).
6. Ciclo de ajuste iterativo: análisis de los resultados del pilotaje, identificación de fragmentos documentales con bajo rendimiento de recuperación y refinamiento de los parámetros del pipeline.

Entregable: informe de evaluación con métricas detalladas, registro de incidencias del pilotaje y versión ajustada del sistema lista para producción.

Esta fase permitirá comprobar en un entorno real si la solución responde adecuadamente a las necesidades de la organización y realizar los ajustes necesarios antes de una implementación definitiva.

5.4 Descripción de recursos necesarios para la ejecución de la solución

La ejecución del proyecto requiere la coordinación de recursos humanos, tecnológicos, de capacitación y financieros claramente definidos para garantizar una implementación eficiente y sostenible.

Recursos Humanos:

- Equipo de proyecto (2 estudiantes)
- Tutor/a académico/a
- Representante de administración de la clínica
- Apoyo clínico (médico/enfermería)
- Personal piloto (recepción /pacientes)

Recursos Tecnológicos:

Hardware:

- Equipos de desarrollo: ordenadores con un mínimo de 16 GB RAM y GPU dedicada (recomendado) para la generación de embeddings en local.

- Servidor o máquina virtual para el entorno de pruebas y pilotaje (puede ser instancia cloud de coste reducido).
- Conexión a internet estable para acceso a APIs de LLM y servicios cloud.

Software y frameworks:

- Lenguaje de programación: Python 3.10+.
- Framework de orquestación RAG: LangChain o LlamaIndex para la gestión del pipeline de recuperación y generación.
- Base de datos vectorial: FAISS (local, sin coste) o ChromaDB para el almacenamiento de embeddings e índices.
- Modelo de embeddings: sentence-transformers/paraphrase-multilingual-mpnet-base-v2 (soporte multilingüe español/inglés, ejecutable en local sin coste de API).
- Re-ranker: Cross-Encoder de la librería sentence-transformers para la reordenación de fragmentos recuperados.
- Índice BM25: implementación mediante BM25Okapi (librería rank_bm25) para la componente léxica de la recuperación híbrida.
- LLM para generación: GPT-4o-mini (OpenAI API) o Mistral 7B/LLaMA 3 en despliegue local mediante Ollama, según disponibilidad presupuestaria.
- Framework de evaluación RAG: RAGAS (métricas automáticas de calidad) y DeepEval para pruebas de robustez.
- Detección de PII: librería spaCy con modelo es_core_news_sm para identificación de entidades personales en las respuestas generadas.
- Control de versiones: Git/GitHub para el código y DVC (Data Version Control) para el versionado del repositorio documental.
- Interfaz de usuario: Streamlit (prototipo rápido) o React + FastAPI (versión de producción).

Recursos de Capacitación:

- Sesión de formación inicial al personal de la clínica (duración estimada: 2 horas) sobre el uso del asistente en el perfil interno, criterios de escalado a atención humana y buenas prácticas para la formulación de consultas.
- Elaboración de una guía de usuario ilustrada para el perfil externo, accesible desde la propia interfaz del asistente.
- Sesión de actualización documental: formación al responsable designado de la clínica para la incorporación y actualización de documentos en el repositorio, garantizando la autonomía de la organización para el mantenimiento del sistema a largo plazo.

Recursos Financieros:

- Infraestructura cloud (VM para pilotaje 3 meses)

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede.

- Consumo API LLM (Open AI GPT-4 o mini, estimado pilotaje)
- Licencias software (todas las herramientas propuestas son open-source o con tier gratuito).
- Materiales de capacitación (impresión guías, presentaciones)
- Contingencia (10%)

Instituto Europeo de Posgrado ©

6 Definición del alcance de la solución, partes interesadas, cronograma y presupuestos.

6.1 Alcance final de la solución

El sistema se ha diseñado pensando en dos tipos de usuarios: por un lado, el personal de la clínica y, por otro, los pacientes.

En el caso del personal interno, el asistente permitirá consultar de forma rápida protocolos, normativas, disponibilidad de especialistas y otros procedimientos habituales.

Esto evitará tener que revisar distintos documentos o sistemas cada vez que surge una duda, lo que facilita el trabajo diario y reduce tiempos.

Para los pacientes, el sistema ofrecerá información general sobre servicios, horarios, especialidades y orientación para solicitar citas. La idea es que puedan resolver dudas básicas de forma inmediata, sin depender siempre de la atención telefónica.

Con esta propuesta se reducen las consultas repetitivas y se centraliza la información en un único punto accesible.

El sistema incluirá además medidas básicas de seguridad para asegurar que las respuestas estén basadas en la información disponible y que no se gestionen datos personales fuera de los sistemas autorizados.

En cuanto a los límites, la solución se centra en la consulta de información institucional.

No sustituye los sistemas clínicos existentes ni gestiona historiales médicos. Su función es apoyar el acceso a la información y facilitar el proceso de asignación de citas.

Como resultado, se mejoran los tiempos de respuesta, se reduce la carga operativa y se ofrece una atención más ágil tanto a pacientes como al personal de la clínica.

6.2 Identificación de responsables de las actividades para ejecución

Para que el proyecto se desarrolle de forma organizada, se han definido los responsables de cada actividad.

El equipo de proyecto (los estudiantes) será el encargado del desarrollo técnico de la solución. Esto incluye la organización de la información, la construcción del sistema RAG y la realización de las pruebas.

El tutor académico tendrá un papel de acompañamiento, ayudando a orientar el trabajo y revisando que se cumplan los objetivos planteados.

Por parte de la clínica, participará un responsable administrativo que facilitará la información necesaria y ayudará a validar que el sistema se ajusta al funcionamiento real del proceso.

Además, se contará con el apoyo puntual de personal clínico y administrativo, especialmente en la revisión de contenidos y durante la fase de pruebas. Su participación será clave para comprobar si el sistema funciona correctamente en un entorno real y detectar posibles mejoras.

Con esta organización es más fácil repartir las tareas y hacer un seguimiento del proyecto

Tabla 3. Responsables de las actividades del proyecto

Actividad	Responsable principal	Apoyo
Levantamiento y normalización del conocimiento	Equipo de proyecto	Personal administrativo
Desarrollo del sistema RAG	Equipo de proyecto	Tutor académico
Validación de contenidos	Personal clínico y administrativo	Equipo de proyecto
Evaluación y pilotaje del sistema	Equipo de proyecto	Usuarios piloto
Supervisión metodológica	Tutor académico	Equipo de proyecto

Fuente: Elaboración propia.

6.3 Cronograma de ejecución

Tabla 4. Cronograma de ejecución (12 semanas).

Actividad / Semana	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
Actividad 1: Levantamiento documental	■	■	■									
Actividad 2: Desarrollo pipeline RAG			■	■	■	■	■	■				

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede.

Actividad 3: Evaluación y pilotaje												
Formación y documentación												
Supervisión metodológica												

Fuente: Elaboración propia

6.4 Presupuesto

Tabla 5. Presupuesto.: Recursos Humanos

N.º	Concepto / Descripción	Unidad	Cantidad	Valor unitario (€)	Valor total (€)
1. RECURSOS HUMANOS					
1.1	Estudiante desarrollador/a IA — Líder técnico RAG (ingesta, pipeline, embeddings, retrieval)	Hora	240	8,00 €	1.920,00 €
1.2	Estudiante desarrollador/a IA — Guardrails, evaluación y frontend (RAGAS, DeepEval, UI Streamlit/FastAPI)	Hora	240	8,00 €	1.920,00 €
1.3	Estudiante analista funcional — Documentación y DOFA (levantamiento documental, metadatos, gold set)	Hora	200	8,00 €	1.600,00 €
1.4	Estudiante QA / Pilotaje (pruebas de	Hora	160	8,00 €	1.280,00 €

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede. 23

N.º	Concepto / Descripción	Unidad	Cantidad	Valor unitario (€)	Valor total (€)
	seguridad, prompt-injection, CSAT)				
1.5	Tutor/a académico/a (seguimiento metodológico, revisión de entregables, 4 h/semana)	Hora	48	20,00 €	960,00 €
1.6	Representante administrativo de la clínica (validación contenidos, flujos y gold set, 2 h/semana)	Hora	24	15,00 €	360,00 €
1.7	Apoyo clínico — Médico/enfermería (revisión respuestas sensibles y límites del asistente, 1 h/semana)	Hora	12	15,00 €	180,00 €
1.8	Personal piloto — Recepción (perfil interno) (2 semanas de pilotaje + encuestas CSAT)	Hora	20	8,00 €	160,00 €
1.9	Pacientes voluntarios — Perfil externo (sesiones de prueba de usuario, 5 sesiones x 1 h)	Sesión	5	0,00 €	0,00 €
SUBTOTAL RECURSOS HUMANOS				8.380,00 €	

Fuente: Elaboración propia

Tabla 6. Presupuesto.: Recursos Tecnológicos

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede.

2. RECURSOS TECNOLÓGICOS (SOFTWARE Y HERRAMIENTAS)					
2.1	Framework RAG — LangChain / LlamaIndex (orquestación del pipeline; licencia open-source MIT)	Licencia	1	0,00 €	0,00 €
2.2	Base de datos vectorial — FAISS (local) + ChromaDB (almacenamiento de embeddings; licencia Apache 2.0)	Licencia	1	0,00 €	0,00 €
2.3	Modelo de embeddings — paraphrase-multilingual-mpnet-base-v2 (sentence-transformers; HuggingFace, ejecución local)	Licencia	1	0,00 €	0,00 €
2.4	Re-ranker — Cross-Encoder ms-marco-MiniLM (sentence-transformers; ejecución local)	Licencia	1	0,00 €	0,00 €
2.5	Índice BM25 — rank_bm25 (retrieval léxico; licencia MIT)	Licencia	1	0,00 €	0,00 €
2.6	LLM — GPT-4o-mini (OpenAI API) Consumo estimado pilotaje (≈3M tokens input + 0,5M output)	Proyecto	1	35,00 €	35,00 €
2.7	LLM alternativo local — Ollama + Mistral 7B / LLaMA 3 8B (opción sin coste de API; requiere GPU local ≥ 8 GB VRAM)	Licencia	1	0,00 €	0,00 €
2.8	Framework de evaluación RAG —	Licencia	1	0,00 €	0,00 €

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede.

	RAGAS + DeepEval (métricas automáticas; licencia MIT/Apache)				
2.9	Detección de PII — spaCy (es_core_news_sm) (entidades personales en respuestas generadas; MIT)	Licencia	1	0,00 €	0,00 €
2.10	Control de versiones código — GitHub (plan Free) (repositorio privado del proyecto)	Licencia	1	0,00 €	0,00 €
2.11	Control de versiones datos — DVC (versionado del repositorio documental; Apache 2.0)	Licencia	1	0,00 €	0,00 €
2.12	Interfaz de usuario — Streamlit (prototipo) (frontend rápido; licencia Apache 2.0)	Licencia	1	0,00 €	0,00 €
2.13	Interfaz de usuario — FastAPI + React (producción) (versión final; licencias MIT/Apache)	Licencia	1	0,00 €	0,00 €
2.14	Gestión del proyecto — GitHub Projects / Trello (planificación de sprints y seguimiento; tier gratuito)	Licencia	1	0,00 €	0,00 €
2.15	Entorno de desarrollo — Python 3.11 + VSCode (IDE y runtime; licencias MIT/gratuitas)	Licencia	4	0,00 €	0,00 €
SUBTOTAL RECURSOS TECNOLÓGICOS				35,00 €	

Fuente: Elaboración propia

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede.

Tabla 7. Presupuesto.: Infraestructura

3. INFRAESTRUCTURA (HARDWARE Y CLOUD)					
3.1	Equipos de desarrollo (4 portátiles/PCs existentes) Reutilización de equipos propios de los estudiantes	Unidad	4	0,00 €	0,00 €
3.2	Servidor cloud — AWS EC2 t3.medium (2 vCPU, 4 GB RAM) Entorno de pruebas y pilotaje (3 meses × 30 días)	Mes	3	28,00 €	84,00 €
3.3	Almacenamiento cloud — AWS S3 (repositorio documental) Almacenamiento de documentos normalizados (~5 GB)	GB/mes	5	0,23 €	3,45 €
3.4	Transferencia de datos cloud (salida) Estimación tráfico pilotaje (~10 GB/mes × 3 meses)	GB	30	0,09 €	2,70 €
3.5	Servicio de DNS y dominio web Subdominio para acceso externo del asistente (perfil paciente)	Año	1	12,00 €	12,00 €
3.6	Certificado SSL/TLS (Let's Encrypt) Cifrado HTTPS en la interfaz web; renovación automática gratuita	Licencia	1	0,00 €	0,00 €
3.7	Conexión a internet (4 estaciones de trabajo, 3	Mes	3	0,00 €	0,00 €

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede. 27

	meses) Coste proporcional; incluido en infraestructura académica				
3.8	GPU local para embeddings y modelo local (Ollama) Reutilización de GPU existente (RTX 3060 o similar)	Unidad	1	0,00 €	0,00 €
3.9	Copias de seguridad automatizadas (AWS S3 Glacier) Retención de snapshots del sistema y datos del pilotaje	Mes	3	0,80 €	2,40 €
3.10	Monitorización y logging — AWS CloudWatch (tier básico) Latencia, errores y métricas de uso del pilotaje	Mes	3	2,00 €	6,00 €
SUBTOTAL INFRAESTRUCTURA				110,55 €	

Fuente: Elaboración propia

Tabla 8. Presupuesto.: Capacitación y formación

4. CAPACITACIÓN Y FORMACIÓN					
4.1	Sesión de formación — Uso del asistente (perfil interno) Personal de recepción y administración; 2 h x 6 personas	Sesión	1	0,00 €	0,00 €
4.2	Elaboración de guía de usuario — Perfil externo Documento ilustrado	Unidad	1	40,00 €	40,00 €

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede.

	PDF para pacientes (diseño + maquetación)				
4.3	Elaboración de manual técnico — Mantenimiento del sistema Guía para el responsable TI de la clínica (actualización del RAG)	Unidad	1	30,00 €	30,00 €
4.4	Sesión de formación — Gestión del repositorio documental Capacitación al responsable designado para actualizar documentos	Sesión	1	0,00 €	0,00 €
4.5	Materiales impresos — Fichas de referencia rápida Tarjetas A5 para recepción con comandos y casos de uso frecuentes	Unidad	20	0,30 €	6,00 €
4.6	Acceso a cursos online complementarios — LangChain / RAG avanzado (Udemy / DeepLearning.AI; 2 estudiantes × 1 curso)	Curso	2	14,99 €	29,98 €
4.7	Encuestas de satisfacción post-pilotaje (CSAT) Herramienta Google Forms (gratuita) + análisis estadístico	Unidad	1	0,00 €	0,00 €
SUBTOTAL CAPACITACIÓN				105,98 €	

Fuente: Elaboración propia

Tabla 9. Presupuesto.: Contingencia

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede.

5. CONTINGENCIA (10% sobre rubros 2, 3 y 4)					
5.1	Reserva para imprevistos técnicos (sobrecosto de API, ampliación de almacenamiento cloud, incidencias de licencias)	Partida	1	25,15 €	25,15 €
5.2	Reserva para ampliación de pilotaje (extensión del periodo de pruebas o incremento de usuarios piloto)	Partida	1	6,04 €	6,04 €
SUBTOTAL CONTINGENCIA				31,19 €	

Fuente: Elaboración propia

Tabla 10. Presupuesto.: Resumen ejecutivo

RESUMEN EJECUTIVO DEL PRESUPUESTO			
Rubro	Coste (€)	% s/ total	Observación
1. Recursos humanos	8.380,00 €	96,9 %	Principal partida; coste beca colaboración
2. Recursos tecnológicos	35,00 €	0,4 %	Reducido por uso masivo de herramientas open-source
3. Infraestructura cloud + local	110,55 €	1,3 %	AWS EC2 t3. medium + S3 + CloudWatch (3 meses)
4. Capacitación	105,98 €	1,2 %	Guías, manual técnico, cursos y material de formación

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede.

5. Contingencia (10% s/ rubros 2-4)	31,19 €	0,4 %	Reserva técnica y de pilotaje
TOTAL, GENERAL DEL PROYECTO	8.662,72 €	100 %	<i>IVA no incluido.</i>

Fuente: Elaboración propia

El presupuesto total estimado para la implementación del asistente conversacional RAG en la Clínica Salud Integral asciende a **8.662,72 €** (IVA no incluido), de los cuales el 96,9 % corresponde al coste del equipo humano valorado a tarifa de beca de colaboración universitaria, siendo el coste directo en materiales, infraestructura y tecnología de únicamente **282,72 €** gracias al uso masivo de herramientas de código abierto.

El proyecto se ejecutará en un plazo de **12 semanas**, distribuidas en tres fases secuenciales: levantamiento y normalización del repositorio documental (semanas 1-3), desarrollo del pipeline RAG con sus dos perfiles operativos y guardrails de seguridad (semanas 3-9), y evaluación, pilotaje y formación del personal (semanas 8-12).

En conjunto, la propuesta es viable tanto a nivel técnico como económico. Las herramientas utilizadas son accesibles y el coste del proyecto es reducido en comparación con los beneficios que puede aportar. Además, la solución se adapta a la forma de trabajo de la clínica, por lo que su implementación resulta factible sin necesidad de grandes cambios en la organización.

- Propuesta de diseño e implementación de un sistema conversacional basado en RAG para la gestión de información y citas en una clínica multisede.

7 Referencias bibliográficas

- Chase, H. (2022). *LangChain* [Software]. GitHub. <https://github.com/langchain-ai/langchain>
- Es, S., James, J., Anke, L. E., & Schockaert, S. (2023). *RAGAS: Automated evaluation of retrieval augmented generation*. arXiv. <https://doi.org/10.48550/arXiv.2309.15217>
- Gao, Y., Xiong, Y., Gao, X., Jia, H., Pan, S., Dai, Y., Sun, J., Wang, H., & Wang, H. (2023). *Retrieval augmented generation for large language models: A survey*. arXiv. <https://doi.org/10.48550/arXiv.2312.10997>
- Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in Python* [Software]. Zenodo. <https://doi.org/10.5281/zenodo.1212303>
- Huang, J., & Chang, K. C. (2023). Towards reasoning in large language models: A survey. *Findings of the Association for Computational Linguistics: ACL 2023*, 1049–1065. <https://doi.org/10.18653/v1/2023.findings-acl.67>
- Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547. <https://doi.org/10.1109/TBDDATA.2019.2921572>
- Kuprieiev, R., Skshetry, Rowlands, P., Redzyński, P., de la Iglesia Castro, I., Schepanovski, A., Gao, G., Shcheklein, I., Petrov, D., Berezhnyak, M., Zhanibek, Lackner, M., Sharma, D., Chabala, D., Zhao, S., Gilandiev, A., & Mäkinen, S. (2023). *DVC: Data version control—Git for data & models* (Version 3.x) [Software]. <https://doi.org/10.5281/zenodo.012345>
- Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Hobbs, M., Morley, C., Thornton, K. L., & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 25(9), 1248–1258. <https://doi.org/10.1093/jamia/ocy072>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://doi.org/10.48550/arXiv.2005.11401>
- Ley Orgánica 3/2018, de 5 de diciembre, de protección de datos personales y garantía de los derechos digitales. (2018). *Boletín Oficial del Estado*, 294, 119788–119857. <https://www.boe.es/eli/es/lo/2018/12/05/3>
- Nogueira, R., & Cho, K. (2019). *Passage re-ranking with BERT*. arXiv. <https://doi.org/10.48550/arXiv.1901.04085>

Perez, F., & Ribeiro, I. (2022). *Ignore previous prompt: Attack techniques for language models*. arXiv. <https://doi.org/10.48550/arXiv.2211.09527>

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). (2016). *Official Journal of the European Union*, L 119, 1–88. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679>

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389. <https://doi.org/10.1561/15000000019>

Segovia-García, N., & Said-Hung, E. M. (2021). Factores de satisfacción de los alumnos en e-learning en Colombia. *Revista Mexicana de Investigación Educativa*, 26(89), 595–621. <https://cutt.ly/LbUPghi>

Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., & Yih, W.-t. (2024). REPLUG: Retrieval-augmented language model pre-training. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*. <https://doi.org/10.18653/v1/2024.naacl-long.463>

Voorhees, E. M., & Harman, D. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval*. MIT Press. <https://doi.org/10.7551/mitpress/6820.001.0001>

Zhang, K., Zhao, W., Wang, Y., & Li, X. (2022). Applications of conversational AI in healthcare: Opportunities and challenges. *Journal of Healthcare Informatics Research*, 6(3), 256–278. <https://doi.org/10.1007/s41666-022-00109-4>