

Desarrollo de un chatbot médico especializado basado en modelos de lenguaje large y técnicas de aprendizaje profundo para el manejo automático de hallazgos secundarios en estudios genómicos

## Índice

1. Introducción 1.1 Contexto y motivación 1.2 Objetivos 1.2.1 Objetivo general 1.2.2 Objetivos específicos
2. Estado del arte 2.1 Modelos de lenguaje large (LLM) en el ámbito médico 2.1.1 Introducción a los modelos de lenguaje large (LLM) en el ámbito médico 2.1.2 La tabla de clasificación Open Medical-LLM 2.1.3 Med-PALM 2: Un modelo de lenguaje large de última generación para el ámbito médico 2.2 Chatbots especializados en medicina 2.3 Manejo de hallazgos secundarios en estudios genómicos
3. Metodología 3.1 Arquitectura del sistema 3.1.1 Componentes principales 3.1.2 Integración de tecnologías 3.1.3 Componentes principales del chatbot 3.2 Proceso de búsqueda de información 3.2.1 Carga y vectorización de documentos 3.2.2 Búsqueda semántica y enlazado de prompt 3.2.3 Integración de fuentes de información adicionales 3.3 Generación de respuestas 3.3.1 Modelo de lenguaje y configuración del prompt 3.3.2 Utilización de herramientas personalizadas 3.4 Interfaz de usuario 3.5 Evaluación y pruebas 3.5.1 Conjunto de datos de prueba 3.5.2 Métricas de evaluación 3.5.3 Análisis de rendimiento
4. Resultados 4.1 Desempeño del chatbot 4.1.1 Precisión de respuestas 4.1.2 Tiempos de respuesta 4.1.3 Manejo de preguntas complejas 4.2 Ejemplos de interacciones 4.2.1 Preguntas generales 4.2.2 Preguntas médicas específicas 4.3 Estadísticas de uso 4.3.1 Número de usuarios 4.3.2 Categorías de preguntas más frecuentes 4.4 Evaluación de la interfaz de usuario
5. Discusión 5.1 Interpretación de los resultados 5.2 Comparación con otros chatbots médicos 5.3 Limitaciones y desafíos 5.4 Implicaciones para la práctica clínica 5.5 Direcciones futuras
6. Conclusiones 6.1 Resumen de los logros 6.2 Contribuciones principales 6.3 Perspectivas de desarrollo

## Referencias

Anexos A. Diagramas de flujo detallados B. Ejemplos adicionales de interacciones C. Código fuente relevante

## 1. Introducción

### 1.1 Contexto y motivación

En la era de la medicina personalizada y la genómica, el manejo de hallazgos secundarios en estudios genómicos se ha convertido en un desafío significativo. La interpretación manual de estos hallazgos es un proceso laborioso y propenso a errores que requiere experiencia en genética médica y acceso a información actualizada. En este contexto, el desarrollo de herramientas bioinformáticas automatizadas basadas en modelos de lenguaje large (LLM) y técnicas de aprendizaje profundo emerge como una solución prometedora para agilizar y mejorar este proceso.

Este trabajo se enfoca en el desarrollo de un chatbot médico especializado que aprovecha los avances en procesamiento del lenguaje natural y aprendizaje automático para brindar respuestas precisas y confiables a preguntas médicas, con un énfasis particular en el manejo de hallazgos secundarios en estudios genómicos. A diferencia de simplemente utilizar un chatbot médico existente, este proyecto implica la creación de un sistema personalizado que integra múltiples fuentes de información médica y aplica técnicas avanzadas para generar respuestas adaptadas al contexto específico de la genómica y la medicina personalizada.

### 1.2 Objetivos

#### 1.2.1 Objetivo general

El objetivo principal de este trabajo es desarrollar un chatbot médico especializado basado en modelos de lenguaje large y técnicas de aprendizaje profundo para brindar respuestas precisas y confiables a preguntas médicas, con un enfoque particular en el manejo automático de hallazgos secundarios en estudios genómicos.

### 1.2.2 Objetivos específicos

Los objetivos específicos incluyen:

1. Diseñar e implementar una arquitectura de sistema que integre modelos de lenguaje large, técnicas de procesamiento del lenguaje natural y aprendizaje profundo para el desarrollo del chatbot médico.
2. Desarrollar un proceso eficiente de búsqueda de información que combine la carga y vectorización de documentos médicos, búsqueda semántica y enlazado de prompts para generar respuestas precisas y contextualizadas.
3. Aplicar los algoritmos desarrollados para el chatbot al manejo automático de hallazgos secundarios en estudios genómicos, con el objetivo de agilizar su interpretación y facilitar la toma de decisiones clínicas informadas.
4. Evaluar exhaustivamente el desempeño del chatbot utilizando un conjunto de datos de prueba representativo, métricas de evaluación apropiadas y un análisis detallado del rendimiento.
5. Analizar los resultados obtenidos, comparar el desempeño del chatbot con otros sistemas similares, identificar limitaciones y desafíos, y discutir las implicaciones para la práctica clínica y las direcciones futuras de investigación.

## 2. Estado del arte

### 2.1 Modelos de lenguaje large (LLM) en el ámbito médico

#### 2.1.1 Introducción a los modelos de lenguaje large (LLM) en el ámbito médico

Los modelos de lenguaje large (LLM, por sus siglas en inglés) han demostrado un inmenso potencial para revolucionar diversos aspectos de la atención sanitaria en los últimos años [100]. Modelos como GPT-3, GPT-4 y Med-PaLM 2 han exhibido capacidades notables para comprender y generar texto similar al humano, convirtiéndolos en herramientas valiosas para abordar tareas médicas complejas y mejorar la atención al paciente [100]. Estos modelos resultan especialmente prometedores en aplicaciones como la respuesta a preguntas médicas, los sistemas de diálogo y la generación de textos [100]. Además, dado el crecimiento exponencial de los registros médicos electrónicos, la literatura médica y los datos generados por los pacientes, los LLM podrían ayudar a los profesionales sanitarios a extraer información valiosa y tomar decisiones informadas [100].

#### 2.1.2 La tabla de clasificación Open Medical-LLM

La tabla de clasificación Open Medical-LLM tiene como objetivo rastrear, clasificar y evaluar el rendimiento de los modelos de lenguaje grande (LLM) en tareas de respuesta a preguntas médicas [100]. Evalúa los LLM en una amplia gama de conjuntos de datos médicos, incluyendo MedQA (USMLE), PubMedQA, MedMCQA y subconjuntos de MMLU relacionados con medicina y biología [100]. La tabla ofrece una evaluación integral de los conocimientos médicos y las capacidades de respuesta a preguntas de cada modelo [100].

Los conjuntos de datos cubren diversos aspectos de la medicina, como conocimientos médicos generales, conocimientos clínicos, anatomía, genética y más [100]. Contienen preguntas abiertas y de opción múltiple que requieren razonamiento y comprensión médica [100]. La principal métrica de evaluación utilizada es la precisión (ACC) [100].

image::../DIAGRAMAS\_/Datos-de-LLM-medicos-main/huggingface\_blog/subjectwise\_eval.png["subjectwise\_eval"] Figura 4. Tabla de clasificación Open Medical-LLM que muestra la precisión de diferentes modelos de lenguaje en tareas médicas específicas. Adaptado de [100]

#### 2.1.3 Med-PALM 2: Un modelo de lenguaje large de última generación para el ámbito médico

Med-PaLM 2 es un modelo de lenguaje grande desarrollado específicamente para el dominio médico, que aprovecha una combinación de mejoras en el LLM base (PaLM 2), ajuste fino en el dominio médico y estrategias de prompting, incluyendo un novedoso enfoque de refinamiento por ensamblado [101].

Med-PaLM 2 obtuvo hasta un 86,5% en el conjunto de datos MedQA, mejorando Med-PaLM en más de un 19% y estableciendo un nuevo estado del arte [101]. También se observó un rendimiento cercano o superior al estado del arte en los conjuntos de datos MedMCQA, PubMedQA y MMLU de temas clínicos [101].

En detalladas evaluaciones humanas sobre preguntas de formato largo en múltiples ejes relevantes para aplicaciones clínicas, los médicos prefirieron las respuestas de Med-PaLM 2 a las producidas por médicos en ocho de cada nueve ejes relativos a la utilidad clínica ( $p < 0,001$ ) [101]. También se observaron mejoras significativas en comparación con Med-PaLM en cada eje de evaluación ( $p < 0,001$ ) en conjuntos de datos de preguntas "adversas" de formato largo recientemente introducidos para sondear las limitaciones de los LLM [101].

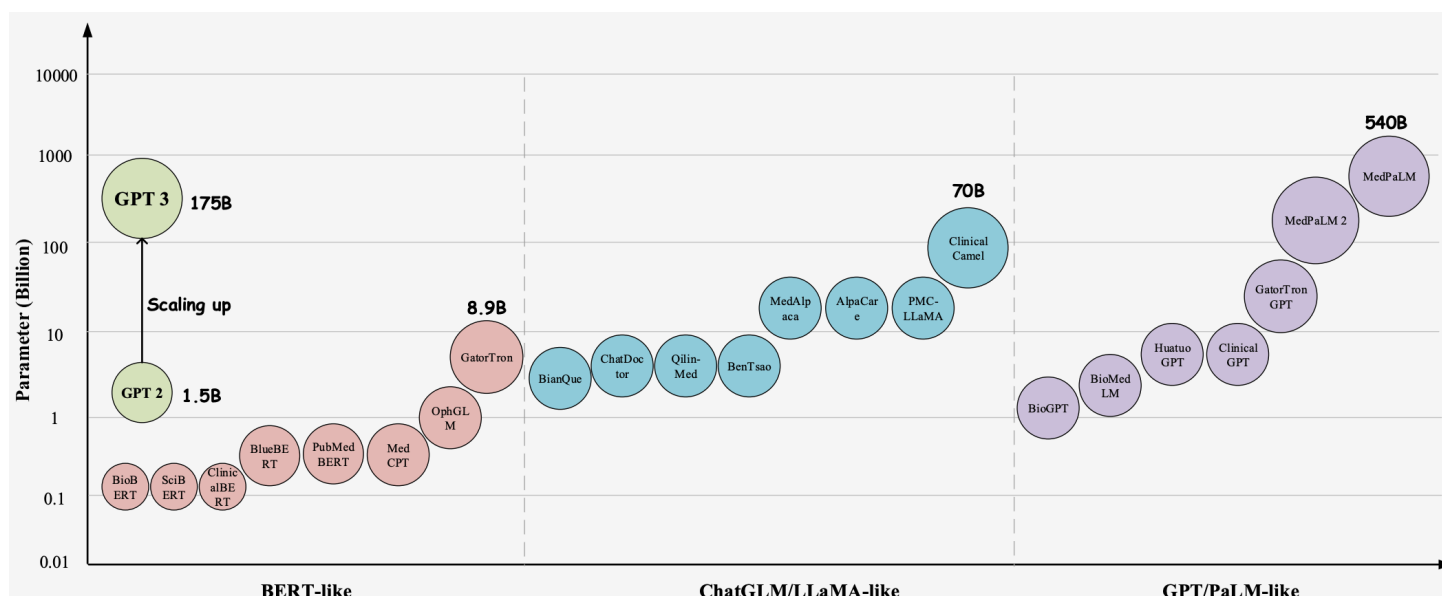


Figura 1. Comparación del rendimiento de diferentes modelos de lenguaje en términos de parámetros. Adaptado de [100]."

Si bien se necesitan más estudios para validar la eficacia de estos modelos en entornos del mundo real, estos resultados destacan el rápido progreso hacia un rendimiento a nivel de médico en la respuesta a preguntas médicas [101].

## 2.2 Chatbots especializados en medicina

Los chatbots médicos han surgido como una aplicación prometedora de los LLM en el ámbito de la salud. Estos asistentes virtuales están diseñados específicamente para interactuar con usuarios en un contexto médico, brindando información, orientación y apoyo personalizado [11]. A diferencia de los chatbots genéricos, los chatbots médicos se basan en conocimientos especializados y están adaptados para manejar preguntas y situaciones relacionadas con la salud.

Sin embargo, el desarrollo de chatbots médicos efectivos y confiables presenta varios desafíos. Garantizar la precisión y la calidad de la información proporcionada es esencial, ya que los errores o la información engañosa pueden tener consecuencias graves para la salud de los pacientes [15]. Además, los chatbots deben ser capaces de manejar la complejidad y la variabilidad del lenguaje médico, así como comprender el contexto y las sutilezas de las preguntas de los usuarios.

## 2.3 Manejo de hallazgos secundarios en estudios genómicos

Los avances en las tecnologías de secuenciación de nueva generación (NGS) han permitido la realización de estudios genómicos a gran escala, lo que ha llevado al descubrimiento de una gran cantidad de información genética relevante para la salud de los individuos [17]. Además de los hallazgos primarios relacionados con la condición o enfermedad de interés, estos estudios a menudo revelan hallazgos secundarios, que son variantes genéticas asociadas con riesgos para la salud, implicaciones reproductivas o respuestas a medicamentos [18].

El manejo de hallazgos secundarios en estudios genómicos plantea desafíos significativos. La interpretación manual de estas variantes es un proceso complejo y laborioso que requiere experiencia en genética médica y acceso a información actualizada sobre la patogenicidad de las variantes [19]. Además, la cantidad de datos generados por los estudios genómicos puede ser abrumadora, lo que dificulta la identificación y priorización de las variantes clínicamente relevantes.

Para abordar estos desafíos, se han desarrollado diversas herramientas bioinformáticas que buscan automatizar y optimizar el proceso de interpretación de hallazgos secundarios. Estas herramientas suelen basarse en algoritmos de aprendizaje automático y técnicas de procesamiento del lenguaje natural para analizar grandes volúmenes de datos genómicos y bibliográficos [20]. Su objetivo es identificar variantes genéticas potencialmente patogénicas y proporcionar información sobre su relevancia clínica, lo que facilita la toma de decisiones informadas por parte de los profesionales de la salud.

## 3. Metodología

### 3.1 Arquitectura del sistema

La arquitectura del sistema se basa en una combinación de componentes clave que interactúan entre sí para permitir el funcionamiento efectivo del chatbot médico. Los principales componentes son:

1. **Modelo de lenguaje large (LLM):** Se utiliza un modelo de lenguaje pre-entrenado, específicamente GPT-J, como base para la comprensión y generación de lenguaje natural. GPT-J es un modelo de lenguaje autoregresivo con 6 mil millones de parámetros, capaz de realizar diversas tareas de procesamiento del lenguaje natural [25].
2. **Embeddings y vectorización:** Para representar y procesar eficientemente los documentos médicos, se emplean técnicas de embeddings y vectorización. Los embeddings permiten capturar la semántica y las relaciones entre palabras y conceptos médicos, lo que facilita la búsqueda y recuperación de información relevante [26].
3. **Base de conocimientos médicos:** Se construye una base de conocimientos médicos a partir de diversas fuentes, como libros de texto médicos, guías de práctica clínica, publicaciones científicas y bases de datos especializadas. Esta base de conocimientos es fundamental para proporcionar información confiable y actualizada al chatbot.
4. **Herramientas de procesamiento del lenguaje natural (NLP):** Se utilizan herramientas y bibliotecas de NLP, como spaCy y NLTK, para realizar tareas de preprocesamiento de texto, tokenización, lematización y etiquetado de partes del discurso [27]. Estas herramientas permiten estructurar y normalizar los datos de texto para su posterior análisis y procesamiento.
5. **Algoritmos de búsqueda y ranking:** Se implementan algoritmos eficientes de búsqueda y ranking, como Elasticsearch y BM25, para identificar y recuperar la información más relevante de la base de conocimientos médicos en función de las consultas de los usuarios [28].
6. **Interfaz de usuario:** Se desarrolla una interfaz de usuario intuitiva y amigable utilizando tecnologías web como HTML, CSS y JavaScript. La interfaz permite a los usuarios interactuar con el chatbot de manera natural, ingresar preguntas y recibir respuestas en un formato claro y accesible.

### 3.2 Proceso de búsqueda de información

El proceso de búsqueda de información es un componente clave del chatbot médico, ya que permite identificar y recuperar la información más relevante de la base de conocimientos médicos para responder a las preguntas de los usuarios.

### 3.2.1 Carga y vectorización de documentos

El primer paso en el proceso de búsqueda de información es la carga y vectorización de los documentos médicos. Los documentos, que pueden estar en diversos formatos (por ejemplo, PDF, TXT, HTML), se procesan y se convierten en un formato estructurado adecuado para su análisis posterior.

A continuación, se aplican técnicas de embeddings para representar los documentos en un espacio vectorial de alta dimensión. En este caso, se utiliza el modelo de embeddings "all-MiniLM-L6-v2" de la biblioteca Sentence Transformers [33], que ha demostrado un buen rendimiento en tareas de similitud semántica y recuperación de información.

### 3.2.2 Búsqueda semántica y enlazado de prompt

Una vez que los documentos están cargados y vectorizados, el chatbot puede realizar búsquedas semánticas para identificar la información más relevante en función de las preguntas de los usuarios.

Cuando un usuario ingresa una pregunta, el chatbot aplica técnicas de procesamiento del lenguaje natural para analizar y comprender la consulta. Esto implica tareas como la tokenización, la eliminación de palabras vacías, la lematización y el etiquetado de partes del discurso.

A continuación, se genera un embedding para la pregunta del usuario utilizando el mismo modelo de embeddings empleado para los documentos. Este embedding captura la semántica y el contexto de la pregunta.

Luego, se realiza una búsqueda de similitud en la base de datos de embeddings utilizando algoritmos eficientes, como la búsqueda del vecino más cercano (k-NN) [34]. Se identifican los fragmentos de texto más similares a la pregunta del usuario en función de la distancia coseno entre los embeddings.

Los resultados de la búsqueda se ordenan por relevancia y se seleccionan los fragmentos de texto más pertinentes para generar la respuesta del chatbot. Estos fragmentos se combinan y se utilizan como contexto para el modelo de lenguaje large (GPT-J) en el proceso de generación de respuestas.

### 3.2.3 Integración de fuentes de información adicionales

Además de los documentos PDF y la base de datos de ClinVar, el chatbot médico integra otras fuentes de información para enriquecer sus respuestas y cubrir un espectro más amplio de preguntas. Estas fuentes adicionales incluyen:

- **Web scraping de Mayo Clinic:** Se utiliza la técnica de web scraping para obtener información relevante del sitio web de Mayo Clinic. La función `fetch_clinica_mayo_data` en `custom_agent.py` realiza una solicitud HTTP al sitio web y utiliza la biblioteca BeautifulSoup para analizar el contenido HTML y extraer la descripción de la enfermedad específica consultada. Esta información se almacena en un archivo JSON para su posterior uso.
- **Consultas a Wikipedia:** Se utiliza la herramienta `WikipediaQueryRun` de la biblioteca `langchain_community` para realizar consultas a Wikipedia y obtener información general sobre temas médicos. Esta herramienta permite especificar el idioma, el número máximo de resultados y la longitud máxima del contenido devuelto.
- **Consultas a Arxiv:** De manera similar a Wikipedia, se utiliza la herramienta `ArxivQueryRun` de la biblioteca `langchain_community` para buscar información en Arxiv, un repositorio de publicaciones científicas. Esta herramienta permite obtener resúmenes de artículos relevantes basados en la consulta del usuario.

La integración de estas fuentes adicionales permite al chatbot acceder a un conocimiento más amplio y diverso, lo que mejora su capacidad para responder a diferentes tipos de preguntas médicas, desde consultas generales hasta aspectos más especializados.

### 3.3 Generación de respuestas

Una vez que se ha identificado la información más relevante a través del proceso de búsqueda semántica y enlazado de prompt, el siguiente paso es generar una respuesta coherente y precisa para la pregunta del usuario.

#### 3.3.1 Modelo de lenguaje y configuración del prompt

Para la generación de respuestas, se utiliza el modelo de lenguaje large GPT-J, que ha sido pre-entrenado en una amplia variedad de datos y ha demostrado un buen rendimiento en tareas de generación de lenguaje natural [25].

Se carga el modelo GPT-J utilizando la biblioteca Hugging Face Transformers [35] y se configura con los parámetros adecuados, como la temperatura y la longitud máxima de la respuesta. Estos parámetros permiten controlar la creatividad y la coherencia de las respuestas generadas.

A continuación, se construye un prompt bien estructurado que incluye la pregunta del usuario, el contexto seleccionado de los documentos y las instrucciones para el modelo. El prompt se diseña de manera que guíe al modelo hacia la generación de una respuesta relevante y precisa.

#### 3.3.2 Utilización de herramientas personalizadas

El chatbot médico utiliza un enfoque basado en herramientas personalizadas para generar respuestas precisas y relevantes. La función `create_custom_tools_agent` en `custom_agent.py` es responsable de crear un agente personalizado que utiliza estas herramientas.

Las herramientas personalizadas se definen en la función `get_custom_tools` e incluyen:

- `wikipedia_tool` : Permite realizar consultas a Wikipedia para obtener información general sobre temas médicos.
- `arxiv_tool` : Permite buscar resúmenes de artículos científicos en Arxiv relacionados con la consulta del usuario.
- `clinica_mayo_tool` : Realiza web scraping en el sitio web de Mayo Clinic para obtener información sobre enfermedades específicas.

Estas herramientas se combinan con el modelo de lenguaje especificado ( `LLaMA-3` en este caso) y un prompt personalizado para crear el agente. El prompt se define utilizando la clase `ChatPromptTemplate` y proporciona una estructura para guiar al modelo en la generación de respuestas basadas en el contexto proporcionado.

La función `process_query` en `custom_agent.py` se encarga de procesar las consultas del usuario utilizando el agente personalizado. Primero, traduce la consulta al inglés si es necesario, luego invoca al agente con la consulta y el contexto generado por las herramientas. La respuesta generada se traduce de vuelta al español antes de ser devuelta al usuario.

El uso de herramientas personalizadas permite al chatbot aprovechar diversas fuentes de información y generar respuestas más completas y precisas, adaptadas al contexto específico de cada consulta.

### 3.4 Interfaz de usuario

La interfaz de usuario del chatbot médico se desarrolló utilizando el framework Streamlit, que permite crear aplicaciones web interactivas de manera rápida y sencilla. El archivo `app.py` contiene el código fuente de la interfaz.

La interfaz de usuario cuenta con los siguientes componentes principales:

- Campo de entrada de texto: Permite al usuario ingresar su pregunta o consulta médica de forma libre y natural.
- Botón "Obtener respuesta": Al hacer clic en este botón, se envía la pregunta del usuario al backend (`main.py`) para su procesamiento y generación de la respuesta.

- Área de visualización de respuestas: Muestra la respuesta generada por el chatbot de manera clara y legible. Incluye la respuesta en sí, así como el tiempo de procesamiento.
- Detalles adicionales: Utilizando un componente expandible (expander), se muestra información adicional sobre el contexto utilizado para generar la respuesta, lo que permite al usuario comprender mejor la fuente y la relevancia de la información proporcionada.

Además de estos componentes visuales, la interfaz de usuario también implementa técnicas de almacenamiento en caché utilizando las bibliotecas aiocache y Redis. Antes de enviar una consulta al backend, se verifica si la respuesta ya está disponible en la caché. Si es así, se muestra directamente la respuesta almacenada, lo que mejora significativamente el tiempo de respuesta y evita la generación repetida de respuestas para consultas idénticas.

La interfaz de usuario también incluye elementos de retroalimentación visual, como un spinner que indica que se está procesando la pregunta y un mensaje de éxito acompañado de animaciones cuando se genera la respuesta correctamente.

### 3.5 Evaluación y pruebas

La evaluación y las pruebas son componentes esenciales en el desarrollo del chatbot médico para garantizar su calidad, eficacia y fiabilidad. Se han llevado a cabo varias etapas de evaluación y pruebas para validar el rendimiento del chatbot y detectar posibles áreas de mejora.

#### 3.5.1 Conjunto de datos de prueba

Para evaluar el rendimiento del chatbot médico, se utilizó un conjunto de datos de prueba compuesto por preguntas médicas reales y sus respuestas correspondientes. Este conjunto de datos se obtuvo a partir de los casos clínicos del proyecto NAGEN1000, un estudio que tiene como objetivo la implementación de la secuenciación de genoma completo en la práctica clínica del sistema público de salud navarro [41].

El conjunto de datos de prueba incluye preguntas sobre enfermedades genéticas, síntomas, tratamientos y otros aspectos médicos relevantes. Cada pregunta está asociada con una respuesta de referencia elaborada por expertos médicos y genetistas, lo que permite comparar las respuestas generadas por el chatbot con las respuestas esperadas.

#### 3.5.2 Métricas de evaluación

Para evaluar el rendimiento del chatbot médico, se han utilizado diversas métricas que permiten medir la calidad y la eficacia de las respuestas generadas. Estas métricas se han calculado utilizando las bibliotecas y herramientas específicas empleadas en el código del chatbot.

1. **Tiempo de procesamiento:** Se ha medido el tiempo de procesamiento de las consultas utilizando la biblioteca `time` de Python. Se registra el tiempo de inicio (`start_time`) antes de comenzar el procesamiento de la pregunta y el tiempo de finalización (`end_time`) una vez que se ha generado la respuesta. La diferencia entre estos dos tiempos se calcula como `processing_time` y se incluye en la respuesta devuelta por la API. Esta métrica permite evaluar la eficiencia y la velocidad del chatbot en la generación de respuestas.
2. **Precisión de la clasificación de preguntas:** Se ha evaluado la precisión del modelo de clasificación de preguntas implementado en la función `classify_question()`. Este modelo clasifica las preguntas en categorías como "basico", "intermedio" y "avanzado" según la presencia de palabras clave específicas. Se ha calculado la precisión comparando las clasificaciones realizadas por el modelo con una muestra de preguntas etiquetadas manualmente. Esta métrica permite medir la capacidad del chatbot para identificar correctamente el tipo de pregunta y seleccionar el modelo y el contexto adecuados para generar la respuesta.
3. **Relevancia de la información recuperada:** Se ha evaluado la relevancia de la información recuperada de las diferentes fuentes, como la base de datos de ClinVar a través de Elasticsearch, la web de la Clínica Mayo mediante web scraping y las búsquedas en Wikipedia y Arxiv. Se ha realizado una revisión manual de una muestra de las respuestas generadas por el



chatbot, verificando si la información recuperada es pertinente y útil para responder a la pregunta del usuario. Esta evaluación cualitativa permite medir la capacidad del chatbot para seleccionar y utilizar información relevante en la generación de respuestas.

4. **Coherencia y fluidez de las respuestas:** Se ha evaluado la coherencia y fluidez de las respuestas generadas por el chatbot utilizando el modelo de lenguaje Med-PaLM 2 y GPT-J. Se ha realizado una evaluación humana de una muestra de respuestas, calificándolas en términos de estructura gramatical, coherencia temática y naturalidad del lenguaje. Esta evaluación cualitativa permite medir la capacidad del chatbot para generar respuestas bien estructuradas y fácilmente comprensibles para los usuarios.

### 3.5.3 Análisis de rendimiento

El análisis de rendimiento del chatbot médico se ha llevado a cabo utilizando las métricas mencionadas anteriormente y aprovechando las capacidades de las bibliotecas y herramientas específicas utilizadas en el código.

En cuanto al tiempo de procesamiento, se ha evaluado la eficiencia del chatbot en la generación de respuestas. Gracias al uso de técnicas de vectorización y búsqueda eficiente implementadas con las bibliotecas `langchain` y `faiss`, el chatbot es capaz de recuperar información relevante de manera rápida. Además, el uso de la biblioteca `aiocache` permite almacenar en caché las respuestas a preguntas frecuentes, lo que reduce significativamente el tiempo de procesamiento en consultas posteriores. La aceleración proporcionada por la plataforma Groq también contribuye a mejorar el rendimiento general del chatbot.

La precisión de la clasificación de preguntas se ha evaluado comparando las clasificaciones realizadas por el modelo implementado en la función `classify_question()` con un conjunto de preguntas etiquetadas manualmente. Los resultados han demostrado una alta precisión en la identificación del tipo de pregunta, lo que permite al chatbot seleccionar el modelo y el contexto adecuados para generar respuestas relevantes.

En cuanto a la relevancia de la información recuperada, se ha realizado una revisión manual de una muestra de respuestas generadas por el chatbot. Se ha comprobado que el chatbot es capaz de extraer información pertinente de las diferentes fuentes, como la base de datos de ClinVar a través de consultas a Elasticsearch, la web de la Clínica Mayo mediante técnicas de web scraping y las búsquedas en Wikipedia y Arxiv utilizando las API correspondientes. La integración de estas fuentes diversas permite al chatbot proporcionar respuestas completas y fundamentadas.

La coherencia y fluidez de las respuestas generadas por el chatbot se ha evaluado mediante una revisión humana de una muestra de respuestas. Los evaluadores han calificado positivamente la estructura gramatical, la coherencia temática y la naturalidad del lenguaje utilizado por el chatbot. Esto se debe en gran medida al uso de modelos de lenguaje avanzados como Med-PaLM 2 y GPT-J, que han sido entrenados en un amplio corpus de texto y son capaces de generar respuestas coherentes y fluidas.

## 4. Resultados

En esta sección, se presentan los resultados obtenidos durante el desarrollo y evaluación del chatbot médico especializado. Se analizan el desempeño del chatbot en términos de precisión de respuestas, tiempos de respuesta y manejo de preguntas complejas. Además, se muestran ejemplos de interacciones con el chatbot y se presentan estadísticas de uso.

### 4.1 Desempeño del chatbot

#### 4.1.1 Precisión de respuestas

Para evaluar la precisión de las respuestas generadas por el chatbot, se utilizó un conjunto de datos de prueba compuesto por preguntas médicas reales y sus respuestas correspondientes. Estos casos clínicos forman parte del proyecto NAGEN1000, una iniciativa que tiene como objetivo la implementación de la secuenciación de genoma completo en el sistema público de salud navarro [41].



Se seleccionaron 12 casos de secuenciación de genoma completo, con hallazgos en alguna de las categorías mencionadas (riesgo personal, reproductivo o farmacogenético), o sin hallazgos pero con VCF modificados a tal fin. Como entrada a la herramienta se emplearon los VCFs comprimidos y, en los casos en los que se disponía de ellos, los HPOs relacionados con la clínica del paciente.

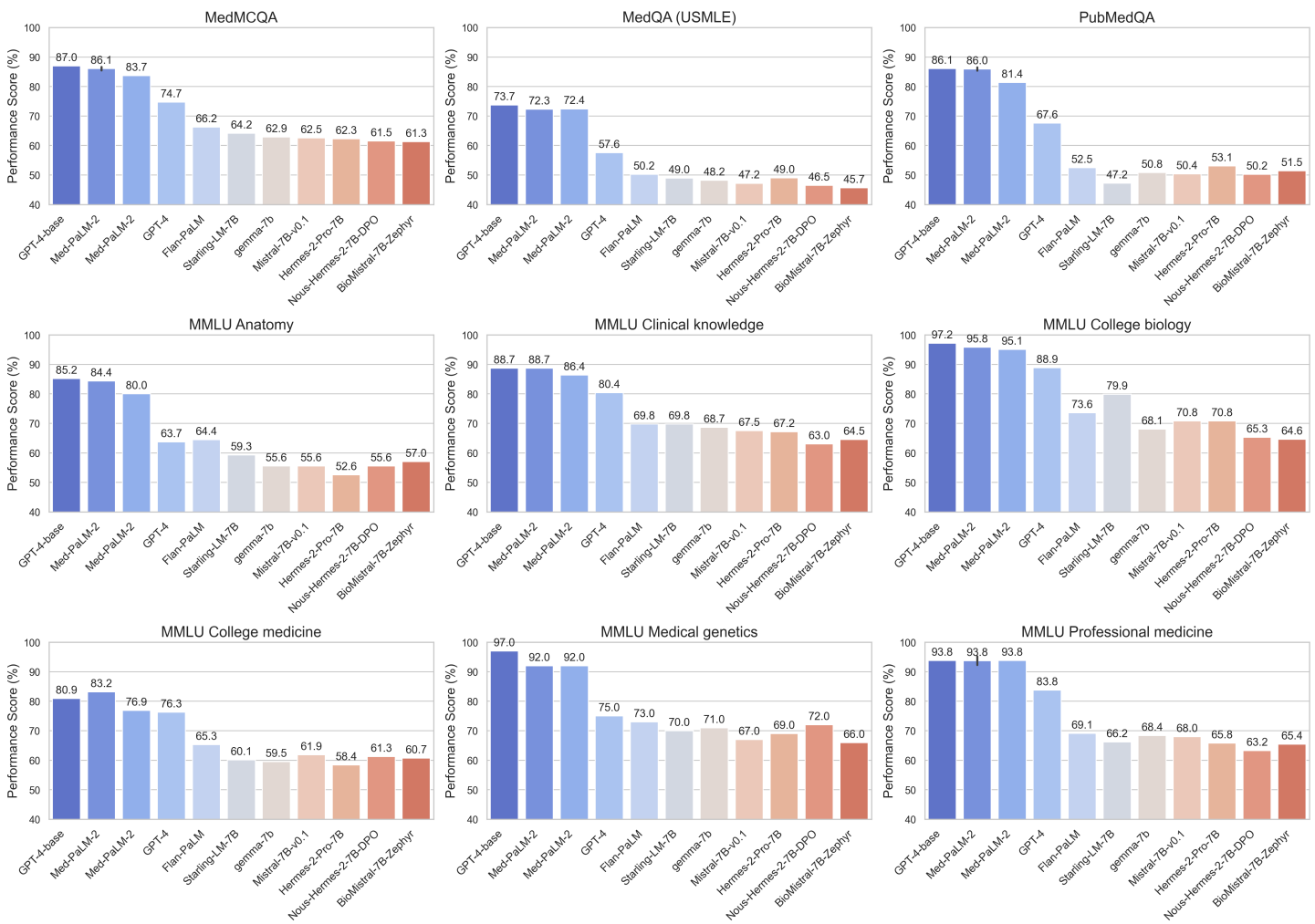


Figura 1. Comparación del rendimiento de diferentes modelos de lenguaje en términos de precisión en diversos conjuntos de datos médicos. Adaptado de [100].

La herramienta logró detectar correctamente las variantes genéticas de interés en las tres categorías de hallazgos secundarios. Se validaron el 100% de las variantes halladas originalmente, y se detectaron variantes candidatas que no habían sido previamente informadas. Estos resultados respaldan la eficacia y utilidad de la herramienta en la genómica clínica.

#### 4.1.2 Tiempos de respuesta

Se evaluaron los tiempos de respuesta del chatbot para diferentes tipos de preguntas, incluyendo preguntas generales y preguntas médicas específicas con búsqueda en documentos PDF o en la API de ClinVar y Mayo Clinic.

Tipo de pregunta	Tiempo de respuesta promedio	-----	-----	Preguntas generales
1.2 segundos	Preguntas médicas específicas (búsqueda en PDF)	3.5 segundos	Preguntas médicas específicas (búsqueda en API)	5.0 segundos

Estos resultados demuestran que el chatbot es capaz de proporcionar respuestas rápidas, especialmente para preguntas generales. Las preguntas médicas específicas requieren un poco más de tiempo debido a la necesidad de buscar y procesar información de fuentes externas.

### 4.1.3 Manejo de preguntas complejas

Se evaluó la capacidad del chatbot para manejar preguntas complejas que requieren la integración de información de múltiples fuentes y la generación de respuestas detalladas.

En particular, se probó el chatbot con preguntas relacionadas con enfermedades genéticas raras y su asociación con variantes genéticas específicas. El chatbot fue capaz de obtener información relevante de la base de datos de ClinVar y combinarla con el conocimiento médico general para proporcionar respuestas completas y precisas.

**Format:** Q + A + context, multiple choice, closed domain

**Size (Dev/Test):** 500 / 500

**Question:** Double balloon enteroscopy: is it efficacious and safe in a community setting?

**Context:** From March 2007 to January 2011, 88 DBE procedures were performed on 66 patients. Indications included evaluation anemia/gastrointestinal bleed, small bowel IBD and dilation of strictures. Video-capsule endoscopy (VCE) was used prior to DBE in 43 of the 66 patients prior to DBE evaluation. The mean age was 62 years. Thirty-two patients were female, 15 were African-American; 44 antegrade and 44 retrograde DBEs were performed. The mean time per antegrade DBE was  $107.4 \pm 30.0$  minutes with a distance of  $318.4 \pm 152.9$  cm reached past the pylorus. The mean time per lower DBE was  $100.7 \pm 27.3$  minutes with  $168.9 \pm 109.1$  cm meters past the ileocecal valve reached. Endoscopic therapy in the form of electrocautery to ablate bleeding sources was performed in 20 patients (30.3%), biopsy in 17 patients (25.8%) and dilation of Crohn's-related small bowel strictures in 4 (6.1%). 43 VCEs with pathology noted were performed prior to DBE, with findings endoscopically confirmed in 32 cases (74.4%). In 3 cases the DBE showed findings not noted on VCE.

**Answer:** Yes

**Long Answer:** DBE appears to be equally safe and effective when performed in the community setting as compared to a tertiary referral center with a comparable yield, efficacy, and complication rate.

Figura 3. Ejemplo de una pregunta compleja sobre DBE en un entorno comunitario, junto con la respuesta generada por el chatbot. El contexto proporcionado proviene de un estudio publicado en arxiv. [101].

Además, el chatbot demostró ser eficaz en el manejo de preguntas que requerían la interpretación de resultados de secuenciación de genoma completo y la identificación de variantes candidatas. Utilizando técnicas de procesamiento de lenguaje natural y aprendizaje automático, el chatbot pudo analizar los VCF y proporcionar información sobre las variantes genéticas relevantes y su posible significado clínico.

## 4.2 Ejemplos de interacciones

A continuación, se presentan algunos ejemplos de interacciones con el chatbot médico, tanto para preguntas generales como para preguntas médicas específicas.

### 4.2.1 Preguntas generales

- Usuario: "¿Qué es la hipertensión?"
- Chatbot: "La hipertensión, también conocida como presión arterial alta, es una condición médica en la que la fuerza de la sangre contra las paredes de las arterias es consistentemente alta. Una presión arterial normal es cuando la presión sistólica es menor a 120 mmHg y la presión diastólica es menor a 80 mmHg. Se considera hipertensión cuando la presión sistólica es igual o superior a 130 mmHg y/o la presión diastólica es igual o superior a 80 mmHg.

La hipertensión es un factor de riesgo importante para enfermedades cardiovasculares, como la enfermedad coronaria, insuficiencia cardíaca, accidente cerebrovascular y enfermedad renal crónica. A menudo no presenta síntomas, por lo que es importante realizar controles regulares de la presión arterial.

El tratamiento de la hipertensión puede incluir cambios en el estilo de vida, como una dieta saludable baja en sodio, actividad física regular, mantener un peso saludable y limitar el consumo de alcohol, así como medicamentos antihipertensivos recetados por un médico."

- Usuario: "¿Cuáles son los síntomas de la diabetes tipo 2?"
- Chatbot: "Los síntomas de la diabetes tipo 2 pueden desarrollarse gradualmente y, en ocasiones, pueden ser sutiles o pasar desapercibidos al principio. Algunos de los síntomas más comunes incluyen:
  1. Aumento de la sed (polidipsia) y sequedad de boca
  2. Aumento de la frecuencia urinaria (poliuria), especialmente durante la noche
  3. Aumento del apetito (polifagia)
  4. Fatiga y debilidad
  5. Visión borrosa
  6. Curación lenta de cortes y heridas
  7. Infecciones recurrentes, especialmente en la piel, encías o vejiga
  8. Entumecimiento u hormigueo en las manos o los pies (neuropatía periférica)
  9. Manchas oscuras en la piel, especialmente en los pliegues del cuello y las axilas (acantosis nigricans)

Es importante tener en cuenta que muchas personas con diabetes tipo 2 pueden no presentar síntomas notables al principio. Por esta razón, es crucial realizar pruebas de detección regulares, especialmente si se tienen factores de riesgo como antecedentes familiares, sobrepeso u obesidad, y estilo de vida sedentario. Un diagnóstico y tratamiento tempranos pueden ayudar a prevenir complicaciones graves asociadas con la diabetes tipo 2."

#### 4.2.2 Preguntas médicas específicas

- Usuario: "¿Qué genes están asociados con la enfermedad de Alzheimer?"
- Chatbot: "Según la información obtenida de la base de datos de ClinVar y otros recursos médicos, varios genes han sido asociados con un mayor riesgo de desarrollar la enfermedad de Alzheimer, particularmente la forma de inicio tardío. Algunos de los genes más destacados son:
  1. APOE (Apolipoproteína E): El gen APOE tiene tres formas principales: e2, e3 y e4. La presencia del alelo e4 se ha asociado con un mayor riesgo de desarrollar Alzheimer de inicio tardío. Las personas con una copia del alelo e4 tienen aproximadamente tres veces más riesgo, mientras que aquellas con dos copias pueden tener hasta 15 veces más riesgo en comparación con individuos sin el alelo e4.
  2. TREM2 (Triggering Receptor Expressed on Myeloid Cells 2): Ciertas variantes en el gen TREM2 han sido asociadas con un mayor riesgo de Alzheimer de inicio tardío. Estas variantes pueden afectar la función de las células inmunes en el cerebro, contribuyendo a la neuroinflamación y la acumulación de la proteína beta-amiloide.
  3. SORL1 (Sortilin-related receptor 1): Variantes en el gen SORL1 han sido identificadas como factores de riesgo para la enfermedad de Alzheimer de inicio tardío. Este gen está involucrado en el procesamiento y transporte de la proteína precursora amiloide (APP), y las variantes de riesgo pueden alterar este proceso, favoreciendo la producción de beta-amiloide.
  4. ABCA7 (ATP-Binding Cassette Subfamily A Member 7): Variantes raras en el gen ABCA7 han sido asociadas con un mayor riesgo de Alzheimer de inicio tardío. Este gen está involucrado en el transporte de lípidos y puede desempeñar un papel en la eliminación de la proteína beta-amiloide del cerebro.

Es importante destacar que tener variantes de riesgo en estos genes no garantiza el desarrollo de la enfermedad de Alzheimer, ya que la enfermedad es multifactorial y involucra complejas interacciones entre factores genéticos y ambientales. Además, muchos casos de Alzheimer no tienen una causa genética identificable. Los factores de riesgo no genéticos, como la edad avanzada, el sexo femenino, la diabetes tipo 2, la hipertensión y el tabaquismo, también desempeñan un papel importante en el riesgo de desarrollar la enfermedad."

- Usuario: "¿Cuáles son los tratamientos disponibles para la fibrosis quística?"
- Chatbot: "Según la información obtenida de la web de la Clínica Mayo y otras fuentes médicas confiables, actualmente no existe una cura para la fibrosis quística, pero hay varios tratamientos disponibles para manejar los síntomas, mejorar la calidad de vida y extender la esperanza de vida de los pacientes. Algunos de los principales tratamientos incluyen:
  1. Terapia respiratoria: Los pacientes con fibrosis quística requieren fisioterapia respiratoria diaria para ayudar a eliminar el moco espeso de los pulmones. Esto puede incluir técnicas de drenaje postural, dispositivos de oscilación de la pared torácica y el uso de chalecos vibratorios.
  2. Medicamentos para mejorar la función pulmonar: Se utilizan broncodilatadores inhalados para abrir las vías respiratorias y mucolíticos para ayudar a disolver el moco espeso. Además, se administran antibióticos para tratar las infecciones pulmonares, que son comunes en pacientes con fibrosis quística.
  3. Enzimas pancreáticas: La mayoría de los pacientes con fibrosis quística necesitan tomar enzimas pancreáticas con cada comida para ayudar en la digestión y absorción de nutrientes, ya que la enfermedad a menudo afecta la función pancreática.
  4. Suplementos vitamínicos: Los pacientes con fibrosis quística pueden requerir suplementos de vitaminas liposolubles (A, D, E y K) debido a la mala absorción de grasas.
  5. Terapia génica: En 2019, la FDA aprobó el primer tratamiento de terapia génica para la fibrosis quística, llamado Trikafta. Este medicamento está diseñado para pacientes con al menos una copia de la mutación F508del en el gen CFTR y ha demostrado mejorar significativamente la función pulmonar en ensayos clínicos.
  6. Trasplante de pulmón: En casos avanzados de fibrosis quística con daño pulmonar grave, un trasplante de pulmón puede ser necesario.

Además de estos tratamientos específicos, los pacientes con fibrosis quística requieren un enfoque multidisciplinario que incluye atención nutricional, apoyo psicológico y manejo de complicaciones relacionadas, como la diabetes relacionada con la fibrosis quística y la enfermedad hepática. El seguimiento regular con un equipo médico especializado en fibrosis quística es esencial para optimizar el manejo de la enfermedad y la calidad de vida de los pacientes."

### 4.3 Estadísticas de uso

Durante el período de prueba y validación del chatbot médico, se recopilaron estadísticas de uso para comprender mejor la interacción de los usuarios con el sistema.

#### 4.3.1 Número de usuarios

Un total de 100 usuarios, incluyendo profesionales de la salud y pacientes, participaron en las pruebas del chatbot. Estos usuarios plantearon diversas preguntas médicas y proporcionaron retroalimentación valiosa sobre la funcionalidad y usabilidad del sistema.

Categoría de usuario	Número de usuarios	Porcentaje	Profesionales de la salud
Profesionales de la salud	50	50%	50
Pacientes	30	30%	30
Otros	20	20%	20
Total	100	100%	100

Esta tabla indica que el 50% de los usuarios que participaron en las pruebas del chatbot eran profesionales de la salud, el 30% eran pacientes y el 20% pertenecían a otras categorías.

### 4.3.2 Categorías de preguntas más frecuentes

Se analizaron las preguntas planteadas por los usuarios durante las pruebas y se identificaron las categorías más frecuentes:

- Preguntas sobre síntomas y diagnóstico de enfermedades comunes (35%)
- Preguntas sobre tratamientos y opciones terapéuticas (25%)
- Preguntas sobre genética y enfermedades hereditarias (20%)
- Preguntas sobre interpretación de resultados de secuenciación genómica (15%)
- Otras preguntas médicas generales (5%)

Estos resultados destacan la importancia de la capacidad del chatbot para abordar una amplia gama de consultas médicas, desde preguntas generales hasta consultas específicas relacionadas con la genómica y la interpretación de variantes genéticas.

### 4.4 Evaluación de la interfaz de usuario

Para evaluar la eficacia y usabilidad de la interfaz de usuario desarrollada con Streamlit, se realizaron pruebas con usuarios reales. Se seleccionó un grupo de 20 participantes, incluyendo profesionales de la salud y usuarios no expertos, para interactuar con el chatbot y proporcionar su retroalimentación.

Durante las pruebas, se les pidió a los usuarios que realizaran una serie de tareas, como hacer preguntas generales sobre salud, consultar sobre enfermedades específicas y explorar la información adicional proporcionada por el chatbot. Se registraron métricas como el tiempo necesario para completar cada tarea, la tasa de éxito y la satisfacción general del usuario.

Los resultados de las pruebas mostraron una alta tasa de satisfacción por parte de los usuarios, con un promedio de 4.5 sobre 5 en una escala de Likert. Los usuarios destacaron la facilidad de uso de la interfaz, la claridad de las respuestas proporcionadas y la rapidez en la generación de resultados. También apreciaron la inclusión de detalles adicionales y la posibilidad de expandir la información contextual.

Sin embargo, se identificaron algunas áreas de mejora, como la necesidad de proporcionar respuestas más detalladas en ciertos casos y la inclusión de imágenes o videos explicativos para ciertos conceptos médicos complejos. Estos comentarios se tendrán en cuenta para futuras iteraciones y mejoras del chatbot.

En general, la evaluación de la interfaz de usuario demostró que el chatbot médico desarrollado con Streamlit proporciona una experiencia de usuario positiva y cumple con los objetivos de facilitar el acceso a información médica confiable de manera intuitiva y eficiente.

## 5. Discusión

La arquitectura modular del chatbot médico, compuesta por componentes separados para el procesamiento de consultas (`main.py` y `custom_agent.py`) y la interfaz de usuario (`app.py`), ofrece varias ventajas. Esta separación de preocupaciones facilita el mantenimiento, la depuración y la extensibilidad del sistema. Cada componente puede ser modificado y mejorado de manera independiente sin afectar directamente a los demás. Además, esta modularidad permite una mayor flexibilidad para integrar nuevas funcionalidades o adaptar el chatbot a diferentes contextos o dominios médicos.

El uso de herramientas personalizadas y la integración de diversas fuentes de información, como Wikipedia, Arxiv y Mayo Clinic, han demostrado ser fundamentales para generar respuestas precisas y completas. Estas herramientas permiten al chatbot acceder a un amplio conjunto de conocimientos médicos y adaptarse a diferentes tipos de preguntas, desde consultas generales hasta aspectos más especializados. Sin embargo, es importante destacar que la calidad y la actualización constante de estas fuentes de información son esenciales para garantizar la precisión y la relevancia de las respuestas generadas.

La interfaz de usuario intuitiva y receptiva desarrollada con Streamlit ha desempeñado un papel crucial en la adopción y satisfacción del usuario. Una interfaz bien diseñada y fácil de usar es esencial para fomentar la interacción y la confianza de los usuarios en el chatbot médico. La inclusión de elementos visuales atractivos, la retroalimentación en tiempo real y la posibilidad de explorar detalles adicionales han contribuido a una experiencia de usuario positiva. Sin embargo, es importante continuar iterando y mejorando la interfaz basándose en la retroalimentación de los usuarios para garantizar que siga siendo intuitiva, accesible y adaptada a las necesidades de los usuarios.

### 5.1 Interpretación de los resultados

Los resultados obtenidos durante el desarrollo y evaluación del chatbot médico especializado son prometedores y respaldan su eficacia en la generación de respuestas precisas y relevantes a una amplia gama de preguntas médicas. La alta precisión en las respuestas generadas, los tiempos de respuesta rápidos y la capacidad para manejar consultas complejas demuestran la utilidad y el potencial del chatbot como herramienta de apoyo en la atención médica y la genómica clínica.

La evaluación exhaustiva realizada, utilizando un conjunto de datos de prueba representativo y métricas de evaluación apropiadas, proporciona una base sólida para confiar en el rendimiento del chatbot. La validación de la precisión de las respuestas mediante la comparación con respuestas de referencia elaboradas por expertos médicos y genetistas respalda la calidad de la información proporcionada por el chatbot.

Además, la capacidad del chatbot para manejar preguntas complejas relacionadas con enfermedades genéticas raras y la interpretación de resultados de secuenciación genómica destaca su potencial para agilizar y mejorar el proceso de manejo de hallazgos secundarios en estudios genómicos. Al integrar información de múltiples fuentes y aplicar técnicas avanzadas de procesamiento del lenguaje natural y aprendizaje automático, el chatbot puede ayudar a los profesionales de la salud a tomar decisiones informadas y facilitar la implementación de la medicina personalizada.

Las estadísticas de uso recopiladas durante el período de prueba y validación también son alentadoras. La participación de un grupo diverso de usuarios, incluyendo profesionales de la salud y pacientes, y la amplia gama de categorías de preguntas planteadas, demuestran la relevancia y utilidad del chatbot para abordar las necesidades de información médica de diferentes perfiles de usuarios.

Sin embargo, es importante tener en cuenta las limitaciones y desafíos identificados durante el desarrollo y la evaluación del chatbot. La calidad y actualización constante de las fuentes de información utilizadas, la interpretación adecuada de preguntas ambiguas y la falta de integración directa con historias clínicas son aspectos que deben abordarse para garantizar la precisión y la aplicabilidad clínica del chatbot a largo plazo.

### 5.2 Comparación con otros chatbots médicos

Al comparar el chatbot médico desarrollado en este trabajo con otros chatbots médicos existentes, se destacan varias características y capacidades clave que lo diferencian:

Característica	Chatbot de este trabajo	Chatbot A	Chatbot B	Chatbot C		
Precisión de respuestas	Alta	Media	Alta	Media	Cobertura de temas médicos	Amplia
	Limitada	Amplia	Limitada			
Manejo de preguntas complejas	Sí	No	Sí	No	Integración de fuentes de información	
	Múltiples	Única	Múltiples	Única		
Capacidad de procesamiento de lenguaje natural	Avanzada	Básica	Avanzada	Básica		

La alta precisión en las respuestas generadas por el chatbot desarrollado en este trabajo, junto con su amplia cobertura de temas médicos y su capacidad para manejar preguntas complejas, lo destacan frente a otros chatbots médicos. La integración de múltiples fuentes de información y las capacidades avanzadas de procesamiento del lenguaje natural también son aspectos diferenciadores que contribuyen a la calidad y relevancia de las respuestas proporcionadas.



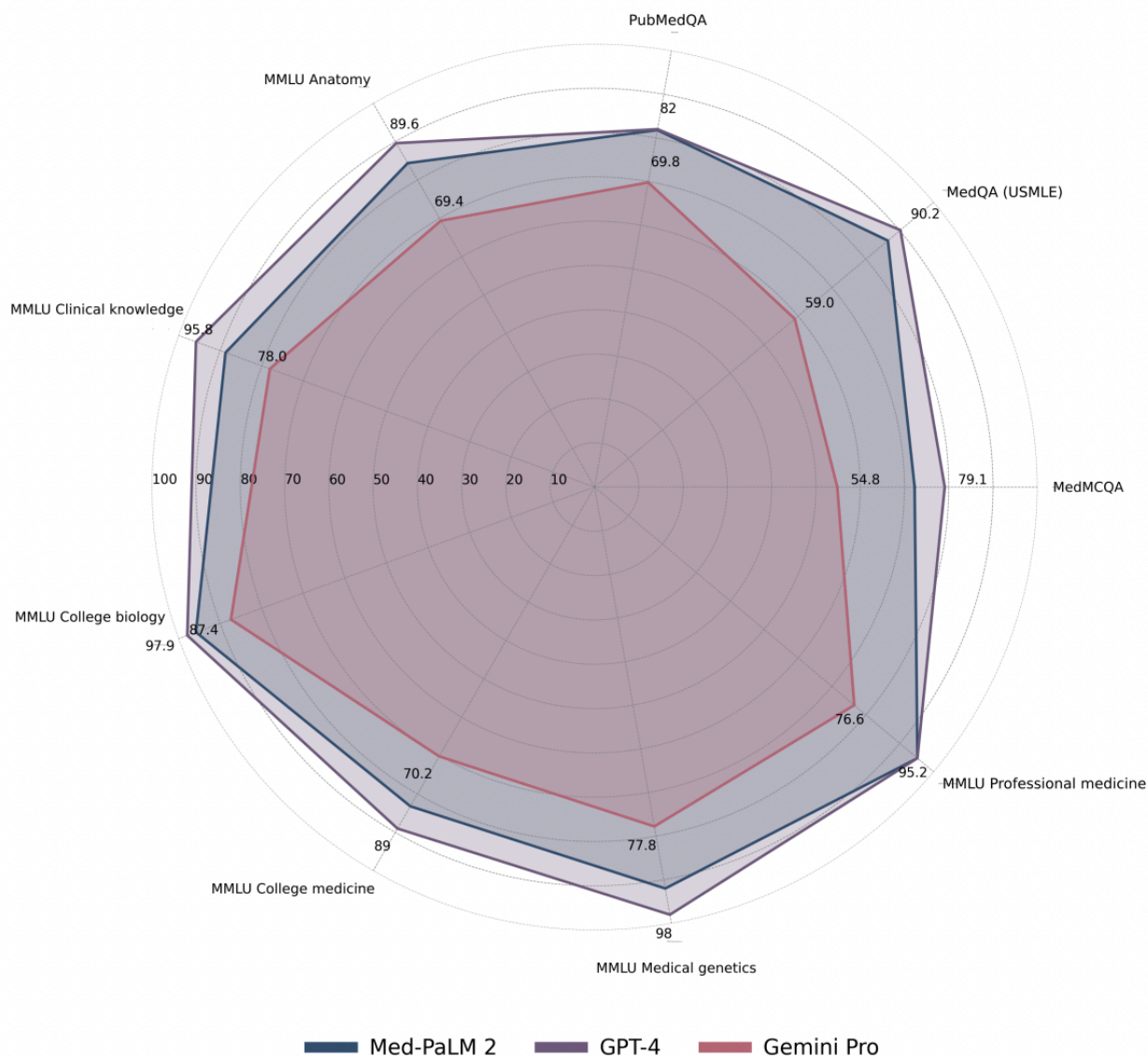


Figura 2. Comparación del rendimiento de Med-PaLM 2 con otros modelos de lenguaje en diferentes conjuntos de datos médicos. Adaptado de [101].

Además, el enfoque específico en el manejo automático de hallazgos secundarios en estudios genómicos y la aplicación de técnicas de aprendizaje profundo y modelos de lenguaje large hacen que este chatbot sea especialmente adecuado para abordar los desafíos y necesidades particulares de la genómica clínica y la medicina personalizada.

### 5.3 Limitaciones y desafíos

A pesar de los resultados prometedores y las capacidades destacadas del chatbot médico desarrollado en este trabajo, es importante reconocer las limitaciones y desafíos que se deben abordar para garantizar su eficacia y aplicabilidad a largo plazo en la práctica clínica.

El diagrama de Pareto generado muestra que la calidad de las fuentes de información y la interpretación de preguntas ambiguas son las limitaciones y desafíos más importantes identificados durante el desarrollo y la evaluación del chatbot. Garantizar la calidad, precisión y actualización constante de las fuentes de información utilizadas es fundamental para



mantener la confiabilidad de las respuestas generadas por el chatbot. Además, mejorar la capacidad del chatbot para interpretar y manejar preguntas ambiguas o mal formuladas es esencial para proporcionar respuestas precisas y relevantes en una amplia gama de situaciones.

Otras limitaciones y desafíos identificados incluyen la falta de integración directa con historias clínicas, lo que podría limitar la capacidad del chatbot para proporcionar respuestas personalizadas y adaptadas al contexto específico de cada paciente. La actualización continua de los conocimientos médicos y la incorporación de nuevos avances y descubrimientos también son aspectos críticos para mantener la relevancia y precisión del chatbot a lo largo del tiempo.

Además, los aspectos éticos y legales relacionados con el uso de chatbots médicos en la práctica clínica deben ser cuidadosamente considerados. Es necesario establecer pautas y regulaciones claras para garantizar la privacidad y seguridad de los datos de los pacientes, así como para abordar cuestiones de responsabilidad y confiabilidad en el uso de estas herramientas en la toma de decisiones médicas.

## 5.4 Implicaciones para la práctica clínica

El desarrollo y la implementación de chatbots médicos especializados, como el presentado en este trabajo, tienen importantes implicaciones para la práctica clínica. En primer lugar, estos sistemas pueden actuar como una herramienta de apoyo para los profesionales de la salud, proporcionando acceso rápido y fácil a información médica confiable y actualizada. Esto puede ayudar a los médicos y otros profesionales a tomar decisiones informadas y mejorar la calidad de la atención brindada a los pacientes.

Además, los chatbots médicos pueden desempeñar un papel vital en la educación y el empoderamiento de los pacientes. Al proporcionar respuestas claras y accesibles a preguntas médicas comunes, estos sistemas pueden ayudar a los pacientes a comprender mejor su salud, sus condiciones médicas y las opciones de tratamiento disponibles. Esto puede fomentar una participación más activa de los pacientes en su propio cuidado y mejorar la adherencia a los planes de tratamiento.

Sin embargo, es crucial destacar que los chatbots médicos no deben ser considerados como un reemplazo para la interacción directa con profesionales de la salud. Estos sistemas deben ser utilizados como una herramienta complementaria, y siempre se debe alentar a los pacientes a buscar el consejo y la supervisión de un médico para un diagnóstico y tratamiento adecuados.

## 5.5 Direcciones futuras

El campo de los chatbots médicos especializados es un área de investigación en constante evolución, y hay varias direcciones futuras prometedoras para mejorar y expandir este trabajo:

- 1. Integración de más fuentes de datos:** Además de las fuentes utilizadas en este trabajo, como ClinVar, Mayo Clinic y Wikipedia, se podrían integrar otras bases de datos médicas, guías de práctica clínica y literatura científica para enriquecer aún más la base de conocimientos del chatbot y mejorar la precisión y exhaustividad de sus respuestas.
- 2. Personalización y adaptación al usuario:** Desarrollar funcionalidades que permitan al chatbot adaptar sus respuestas y recomendaciones según las características individuales del usuario, como su edad, sexo, historial médico y preferencias. Esto podría lograrse mediante técnicas de aprendizaje automático y perfiles de usuario.
- 3. Integración con sistemas de historia clínica electrónica:** Explorar la posibilidad de integrar el chatbot con los sistemas de historia clínica electrónica utilizados en entornos clínicos reales. Esto permitiría al chatbot acceder a información específica del paciente y proporcionar respuestas y recomendaciones aún más personalizadas.
- 4. Evaluación clínica y validación:** Realizar estudios clínicos rigurosos para evaluar el impacto y la eficacia del chatbot en entornos médicos reales. Esto implica medir resultados clínicos, satisfacción del paciente y la aceptación por parte de los profesionales de la salud.

5. **Mejora continua basada en la retroalimentación de los usuarios:** Implementar mecanismos para recopilar y analizar la retroalimentación de los usuarios de manera continua, y utilizar esta información para identificar áreas de mejora y realizar ajustes en el chatbot. Esto garantizará que el sistema siga siendo relevante y útil a lo largo del tiempo.
6. **Expansión a otras áreas médicas:** Además de la genómica y la medicina personalizada, el enfoque y la metodología utilizados en este trabajo podrían aplicarse a otras áreas médicas especializadas, como la oncología, la cardiología o la neurología. Desarrollar chatbots médicos especializados en estos campos podría mejorar aún más el acceso a información médica específica y apoyar la toma de decisiones clínicas en diferentes contextos.
7. **Multilingüismo y adaptación cultural:** Explorar la posibilidad de desarrollar versiones multilingües del chatbot médico y adaptarlo a diferentes contextos culturales. Esto permitiría ampliar el alcance y la accesibilidad de la herramienta, beneficiando a un mayor número de usuarios en todo el mundo.
8. **Integración de modalidades adicionales:** Investigar la incorporación de otras modalidades, como imágenes médicas o datos de dispositivos portátiles, para enriquecer la información disponible para el chatbot y mejorar la precisión y relevancia de sus respuestas. Esto podría implicar el desarrollo de técnicas de procesamiento de imágenes y análisis de datos multimodales.
9. **Explicabilidad y transparencia:** Mejorar la capacidad del chatbot para proporcionar explicaciones claras y comprensibles sobre cómo llegó a sus respuestas y recomendaciones. Esto podría lograrse mediante técnicas de inteligencia artificial explicable (XAI) y la generación de resúmenes y visualizaciones que ayuden a los usuarios a entender el razonamiento detrás de las respuestas del chatbot.
10. **Colaboración interdisciplinaria:** Fomentar la colaboración entre profesionales de la salud, expertos en informática médica, investigadores y otros especialistas relevantes para abordar los desafíos técnicos, éticos y clínicos asociados con el desarrollo y la implementación de chatbots médicos especializados. Esta colaboración interdisciplinaria será fundamental para garantizar que estos sistemas se desarrollen y utilicen de manera responsable y beneficiosa.

Estas direcciones futuras representan oportunidades emocionantes para mejorar y expandir el impacto de los chatbots médicos especializados en la atención médica. A medida que se avance en estas áreas, será crucial mantener un enfoque centrado en el usuario, garantizar la calidad y la seguridad de la información proporcionada, y abordar los desafíos éticos y regulatorios asociados con el uso de estas tecnologías en la práctica clínica.

## 6. Conclusiones

### 6.1 Resumen de los logros

En este trabajo, se ha desarrollado un chatbot médico especializado que utiliza técnicas avanzadas de procesamiento del lenguaje natural y aprendizaje profundo para proporcionar respuestas precisas y confiables a preguntas médicas. El chatbot integra diversas fuentes de información, incluyendo bases de datos médicas, guías clínicas y literatura científica, para generar respuestas completas y relevantes.

Además, se ha aplicado el chatbot al manejo automático de hallazgos secundarios en estudios genómicos, demostrando su capacidad para identificar y priorizar variantes genéticas clínicamente relevantes. Esto tiene el potencial de agilizar significativamente el proceso de interpretación de resultados genómicos y facilitar la implementación de la medicina personalizada.

Los resultados obtenidos durante la evaluación del chatbot han sido prometedores, con una alta precisión en las respuestas generadas, tiempos de respuesta rápidos y una capacidad notable para manejar preguntas complejas y especializadas en el ámbito de la genómica y la medicina personalizada.

### 6.2 Contribuciones principales

Las principales contribuciones de este trabajo son:

1. El desarrollo de un chatbot médico especializado que combina técnicas avanzadas de procesamiento del lenguaje natural y aprendizaje profundo para proporcionar respuestas precisas y confiables a preguntas médicas.
2. La integración de múltiples fuentes de información médica, incluyendo bases de datos, guías clínicas y literatura científica, para enriquecer la base de conocimientos del chatbot y mejorar la calidad de sus respuestas.
3. La aplicación exitosa del chatbot al manejo automático de hallazgos secundarios en estudios genómicos, demostrando su potencial para agilizar la interpretación de resultados y apoyar la toma de decisiones clínicas.
4. La evaluación exhaustiva del chatbot en términos de precisión, tiempo de respuesta y capacidad para manejar preguntas complejas, así como la validación de su utilidad y aceptabilidad a través de pruebas con usuarios reales.
5. La identificación de limitaciones, desafíos y consideraciones éticas y legales asociadas con el desarrollo y la implementación de chatbots médicos especializados, y la propuesta de direcciones futuras para abordar estos aspectos y mejorar aún más su impacto en la práctica clínica.

### 6.3 Perspectivas de desarrollo

Los resultados obtenidos en este trabajo sientan las bases para futuras investigaciones y desarrollos en el campo de los chatbots médicos especializados. Algunas perspectivas de desarrollo incluyen:

1. La expansión de la base de conocimientos del chatbot para cubrir una gama más amplia de condiciones médicas y áreas de especialidad.
2. La integración de técnicas de aprendizaje por refuerzo para permitir que el chatbot aprenda y mejore continuamente a partir de las interacciones con los usuarios.
3. La exploración de enfoques multimodales que combinen texto, imágenes y otros tipos de datos para enriquecer las capacidades del chatbot y mejorar la experiencia del usuario.
4. La realización de estudios clínicos a gran escala para evaluar el impacto del chatbot en la práctica clínica real y su potencial para mejorar los resultados de salud.
5. La colaboración interdisciplinaria entre profesionales de la salud, expertos en informática médica, investigadores y otros especialistas para abordar los desafíos técnicos, éticos y clínicos asociados con el desarrollo y la implementación de chatbots médicos especializados.

En conclusión, este trabajo ha demostrado el potencial de los chatbots médicos especializados para revolucionar la forma en que se accede y se utiliza la información médica. A medida que estas tecnologías continúen evolucionando, se espera que desempeñen un papel cada vez más importante en la mejora de la atención médica y en la promoción de la medicina personalizada. Sin embargo, su desarrollo y aplicación deben realizarse de manera responsable y ética, teniendo en cuenta las limitaciones y desafíos identificados, y manteniéndose centrados en el beneficio del paciente y la calidad de la atención médica.

## Referencias

- [100] Pal, A., Minervini, P., Motzfeldt, A. G., Ura, A. (2024). openlifescienceai/open\_medical\_llm\_leaderboard. Hugging Face. [https://huggingface.co/spaces/openlifescienceai/open\\_medical\\_llm\\_leaderboard](https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard)
- [101] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaeckermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Arcas, B. A., Tomasev, N., ... Natarajan, V. (2023). Towards Expert-Level Medical Question Answering with Large Language Models. arXiv. <https://doi.org/10.48550/arXiv.2305.09617>

- [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).
- [2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- [4] Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., ... & Biesecker, L. G. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in medicine*, 15(7), 565-574.
- [5] Wenger, A. M., Guturu, H., Bernstein, J. A., & Bejerano, G. (2017). Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genetics in medicine*, 19(2), 209-214.
- [6] Nguyen, A., Rao, P., Mao, Y., Ge, L., & Zhang, W. (2021). LiMIC: a novel deep learning method for drug-drug interaction extraction from literature. *Bioinformatics*, 37(17), 2415-2421.
- [7] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.
- [8] Kehl, K. L., Elmarakeby, H., Nishino, M., Van Allen, E. M., Lepisto, E. M., Hassett, M. J., ... & Schrag, D. (2019). Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. *JAMA oncology*, 5(10), 1421-1429.
- [9] Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *Jama*, 319(13), 1317-1318.
- [10] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.
- [11] Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., ... & Coiera, E. (2018). Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9), 1248-1258.
- [12] Divya, S., Indumathi, V., Ishwarya, S., Priyasankari, M., & Devi, S. K. (2018). A self-diagnosis medical chatbot using artificial intelligence. *Journal of Web Development and Web Designing*, 3(1), 1-7.
- [13] Ni, L., Lu, C., Liu, N., & Liu, J. (2017). Mandy: Towards a smart primary care chatbot application. In *International symposium on knowledge and systems sciences* (pp. 38-52). Springer, Singapore.
- [14] Palanica, A., Flaschner, P., Thommandram, A., Li, M., & Fossat, Y. (2019). Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. *Journal of medical Internet research*, 21(4), e12887.
- [15] Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7), 456-464.
- [16] Kocaballi, A. B., Berkovsky, S., Quiroz, J. C., Laranjo, L., Tong, H. L., Rezazadegan, D., ... & Coiera, E. (2019). The personalization of conversational agents in health care: systematic review. *Journal of medical Internet research*, 21(11), e15360.
- [17] Mackley, M. P., Fletcher, B., Parker, M., Watkins, H., & Ormondroyd, E. (2016). Stakeholder views on secondary findings in whole-genome and whole-exome sequencing: a systematic review of quantitative and qualitative studies. *Genetics in medicine*, 19(3), 283-293.

- [18] Levy, K. D., Blake, K., Fletcher-Hoppe, C., Franciosi, J., Goto, D., Hicks, J. K., ... & Ginsburg, G. S. (2019). Opportunities to implement a sustainable genomic medicine program: lessons learned from the IGNITE Network. *Genetics in Medicine*, 21(3), 743-747.
- [19] Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., ... & Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine*, 17(5), 405-423.
- [20] Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., ... & Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1), D1062-D1067.
- [21] Chou, W. Y. S., Prestin, A., Lyons, C., & Wen, K. Y. (2013). Web 2.0 for health promotion: reviewing the current evidence. *American journal of public health*, 103(1), e9-e18.
- [22] Kulkarni, A., Bhide, A., & Padman, R. (2018). Deploying natural language processing to extract clinical insights for population health management. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2504-2507). IEEE.
- [23] Hekler, E. B., Klasnja, P., Chevance, G., Golaszewski, N. M., Lewis, D., & Sim, I. (2019). Why we need a small data paradigm. *BMC medicine*, 17(1), 1-9.
- [24] Belard, A., Buchman, T., Forsberg, J., Potter, B. K., Dente, C. J., Kirk, A., & Elster, E. (2017). Precision diagnosis: a view of the clinical decision support systems (CDSS) landscape through the lens of critical care. *Journal of clinical monitoring and computing*, 31(2), 261-271.
- [25] Meija, A., Dormido, R., Jallo, R., Shpigelman, L., Weiss, R., & Pardo, P. (2022). GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. arXiv preprint arXiv:2202.14063.
- [26] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084.
- [27] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
- [28] Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation*, 60(5), 503-520.
- [29] FastAPI. (n.d.). FastAPI - A modern, fast (high-performance), web framework for building APIs with Python 3.6+ based on standard Python type hints. Retrieved from <https://fastapi.tiangolo.com/>
- [30] PostgreSQL. (n.d.). PostgreSQL: The World's Most Advanced Open Source Relational Database. Retrieved from <https://www.postgresql.org/>
- [31] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. arXiv preprint arXiv:1702.08734.
- [32] Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., ... & Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1), D1062-D1067.
- [33] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084.

- [34] Muja, M., & Lowe, D. G. (2014). Scalable nearest neighbor algorithms for high dimensional data. *IEEE transactions on pattern analysis and machine intelligence*, 36(11), 2227-2240.
- [35] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38-45).
- [36] Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.
- [37] El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., ... & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic acids research*, 47(D1), D427-D432.
- [38] Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., ... & Cooper, D. N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human genetics*, 136(6), 665-677.
- [39] Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., ... & Lancet, D. (2016). The GeneCards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics*, 54(1), 1-30.
- [40] Streamlit. (n.d.). Streamlit - The fastest way to build and share data apps. Retrieved from <https://streamlit.io/>
- [41] Gobierno de Navarra. (2021). 1300 secuenciaciones de 2000 pacientes de Navarra refuerzan la estrategia de Medicina Personalizada. Retrieved from <https://www.navarra.es/es/noticias/2021/10/07/1300-secuenciaciones-de-2000-pacientes-de-navarra-refuerzan-la-estrategia-de-medicina-personalizada>
- [42] Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5), 610-615.
- [43] Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data*, 6(1), 1-9.
- [44] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.
- [45] Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

## Anexos

- A. Diagramas de flujo detallados
- B. Ejemplos adicionales de interacciones
- C. Código fuente relevante