



UNIVERSIDAD DE GRANADA

Escuela Técnica Superior de Ingenierías Informática y de
Telecomunicación

GRADO EN INGENIERÍA INFORMÁTICA

TRABAJO DE FIN DE GRADO

Aplicación de Ayuda para la Generación de Informes Clínicos

Presentado por:
Raúl Martínez Alonso

Curso académico 2024-2025

Aplicación de Ayuda para la Generación de Informes Clínicos

Raúl Martínez Alonso

Raúl Martínez Alonso *Aplicación de Ayuda para la Generación de Informes Clínicos.*
Trabajo de fin de Grado. Curso académico 2024-2025.

**Responsable de
tutorización**

Carlos Fernández Basso
*Departamento de Lenguajes y Sistemas
Informáticos*

Andrea Morales Garzón
*Departamento de Ciencias de la Computación
e Inteligencia Artificial*

Grado en Ingeniería
Informática

Escuela Técnica Superior
de Ingenierías Informática
y de Telecomunicación
Universidad de Granada

DECLARACIÓN DE ORIGINALIDAD

D./Dña. Raúl Martínez Alonso

Declaro explícitamente que el trabajo presentado como Trabajo de Fin de Grado (TFG), correspondiente al curso académico 2024-2025, es original, entendido esto en el sentido de que no he utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a 16 de junio de 2025

A handwritten signature in black ink, appearing to read "RM".

Fdo: Raúl Martínez Alonso

Agradecimientos

A mis padres, abuelos y hermana, por haberme apoyado siempre en cada paso que doy. Sin vuestra ayuda, esto no hubiera sido posible.

Gracias a mis padres por siempre haberme cuidado, educado y estado pendientes de mí en cada momento. Gracias a mi abuela Pilar por siempre sacarme una sonrisa.

Por último, gracias a mis amigos, a los de siempre y a los que he conocido en esta etapa, por todo el apoyo.

Resumen

Palabras clave: Artificial Intelligence, Natural Language Processing, Text Generation, Generative Models, Clinical Reports, MIMIC-IV

Este proyecto tiene como objetivo desarrollar una aplicación capaz de generar automáticamente informes clínicos a partir de textos o diagnósticos médicos, utilizando modelos de Inteligencia Artificial (IA) generativa. El objetivo principal es aliviar la carga administrativa de los profesionales sanitarios, reduciendo tanto el tiempo como el esfuerzo necesarios para redactar informes, al mismo tiempo que se mejora la consistencia, la coherencia y la precisión de la información recogida en ellos.

Los avances recientes en el procesamiento de lenguaje natural (NLP), en particular con modelos generativos de gran tamaño, han abierto nuevas alternativas para automatizar la producción de informes clínicos. Modelos de lenguaje de última generación (LLM), como GPT, LLaMA, Qwen o Phi, han demostrado una notable capacidad para producir textos coherentes y contextualmente adecuados cuando reciben una instrucción o “prompt” específico.

Este proyecto utiliza el conjunto de datos MIMIC-IV, una colección de información clínica de pacientes de acceso público. El procedimiento metodológico comienza con el procesado y la limpieza de los datos, eliminando redundancias y conservando los informes clínicos más relevantes. Posteriormente, los datos preparados se utilizan como contexto para que los modelos generativos puedan producir informes clínicos más concisos pero completos a partir de determinados parámetros de entrada.

La principal funcionalidad de la aplicación consiste en resumir textos clínicos extensos y generar así informes más ágiles y estructurados. Los modelos se han evaluado utilizando métricas automáticas, como BLEU, que mide el solapamiento y la similitud entre el informe original y el texto resumido. Además, se han realizado diferentes pruebas para determinar la influencia de añadir un rol especializado al prompt y de modificar la temperatura en el procedimiento de generación.

Los resultados muestran que la aplicación logra producir informes clínicos coherentes, clínicamente más concisos pero sin perder información relevante. Esto pone de manifiesto el potencial de la Inteligencia Artificial para aliviar la carga documental en el entorno sanitario, liberando así más tiempo y recursos de los trabajadores de la salud para atender directamente a los pacientes. Además, evidencia cómo conjuntos de datos específicos como MIMIC-IV pueden utilizarse para adaptar modelos de IA a aplicaciones muy concretas y de alto impacto en el sector médico.

Summary

Keywords: Artificial Intelligence, Natural Language Processing, Text Generation, Generative Models, Clinical Reports, MIMIC-IV.

This project aims to develop an application that can automatically generate clinical reports from medical texts or diagnoses using generative Artificial Intelligence (AI) models. The main objective is to ease the administrative burden on healthcare professionals by reducing the time and effort required to produce reports, while improving their consistency, coherence, and accuracy.

Recent advances in Natural Language Processing (NLP), particularly with large generative models, have opened up new possibilities for automating the production of medical reports. Large Language Models (LLMs) such as GPT, LLaMA, Qwen, or Phi have demonstrated a remarkable ability to generate coherent and contextually appropriate text when conditioned on a prompt.

This project utilizes the MIMIC-IV dataset, a large, publicly available collection of patients data. The first step in the methodology involves processing and cleaning the data, removing redundancy, and retaining the clinical reports. Subsequently, the data is used as context for generative models to produce accurate and coherent reports from a given prompt.

The main functionality of the application comprises summarizing extensive clinical texts and generating a well-structured report. The models were evaluated using automatic metrics, such as BLEU, which measures the overlap and similarity between the generated reports and the original documents. Additionally, different experiments were conducted to assess the influence of adding a specialized role to the prompt and modifying temperature.

The results demonstrate that the application successfully produces coherent and clinically accurate reports, reducing their length while retaining essential information.

This approach underscores the potential of Artificial Intelligence to ease the administrative load on healthcare workers, allowing them to devote more time and resources to patient care. It also shows how large, specialized datasets like MIMIC-IV can be leveraged to fine-tune models for specific, high-impact applications in medicine.

Índice general

Agradecimientos	III
Resumen	v
Summary	VII
1. Introducción	1
1.1. Contexto y motivación	1
1.2. Problema a resolver	2
1.3. Objetivos	2
1.3.1. Objetivo general del trabajo	2
1.3.2. Objetivos específicos	3
1.4. Planificación y costes del proyecto	3
1.4.1. Planificación del proyecto	3
1.4.2. Costes del proyecto	4
1.5. Estructura de la memoria	5
2. Estado del arte	7
2.1. Inteligencia artificial en salud	7
2.2. Modelos de IA para generación automática de texto	7
2.3. Generación de informes clínicos usando IA generativa	9
2.4. Uso del dataset MIMIC-IV en la investigación	9
3. Metodología	11
3.1. Unificación y preprocesamiento de datos clínicos	11
3.2. Generación automática de informes clínicos	11
3.3. Justificación de los modelos de lenguaje utilizados	12
4. Conjunto de datos, descripción y unificación	13
4.1. Análisis y descripción del dataset	13
4.1.1. Anonimización de datos	15
4.2. Preprocesamiento y unificación	17
4.2.1. Preprocesamiento módulo hosp	18
4.2.2. Preprocesamiento módulo icu	18
4.2.3. Preprocesamiento módulo notes	19
4.2.4. Filtración subject_id presentes en módulo notes	19
4.2.5. Concatenación de todos los subconjuntos	19
4.3. Análisis exploratorio de datos	20
4.3.1. Distribución de la edad de los pacientes	21
4.3.2. Ubicación de alta del paciente	21
4.3.3. Distribución por género	21
4.3.4. Enfermedades más frecuentes	22
4.3.5. Análisis individual de un paciente	23

Índice general

4.3.6. Diagnóstico de un paciente	23
4.3.7. Wordcloud	23
5. Generación automática de informes clínicos	29
5.1. Procedimiento	29
5.2. Métricas de evaluación	29
5.2.1. Métrica de calidad de generación del texto: BLEU	30
5.2.2. Métrica de comparación de similitud de las respuestas: Bert score	31
5.2.3. Medidas de readabilidad y comprensión del texto	31
6. Experimentos y resultados	33
6.1. Diseño de la experimentación	33
6.1.1. Resumen de consulta clínica	33
6.1.2. Prueba de ventana de contexto	34
6.1.3. Justificación de la elección de los modelos	34
6.1.4. Justificación de la elección de los conjuntos de datos	35
6.2. Tecnologías utilizadas	36
6.3. Análisis de métricas	37
6.3.1. Longitud media de los textos	37
6.3.2. Métrica BLEU	38
6.3.3. Métrica de comparación de similitud de las respuestas: Bert score	39
6.3.4. Medidas de readabilidad y comprensión del texto	41
6.3.5. Consideraciones finales	44
7. Conclusiones y trabajo futuro	49
7.1. Conclusiones	49
7.2. Trabajos futuros	50
A. Apéndice	53
A.1. Ejemplo de diagnóstico de un paciente	53
A.2. Ejemplo de la primera experimentación: resumen de consulta clínica	58
A.3. Ejemplo de la segunda experimentación: prueba de ventana de contexto	63
Bibliografía	67

1. Introducción

En este primer capítulo se expone de manera desarrollada el contexto y motivación del proyecto, así como el problema a resolver y los objetivos planteados para ello. A continuación, se describe la planificación y costes del proyecto. Finalmente, se aclara la estructura de la memoria.

1.1. Contexto y motivación

En el contexto de la atención médica, los informes clínicos son esenciales no solo para el seguimiento de la salud del paciente, sino que también recogen diagnósticos, comunicación entre profesionales y toma de decisiones. Estos documentos, elaborados por profesionales sanitarios, recogen información relevante de manera estructurada y detallada. Suelen contener antecedentes, síntomas, pruebas médicas, tratamientos y evolución de los pacientes, entre otros aspectos relevantes. Su correcta elaboración es crucial tanto desde el punto de vista médico como legal.

No obstante, la redacción de estos informes supone una carga considerable para los sanitarios, reduciendo la eficiencia de los servicios de salud y restando tiempo a la atención del paciente. En la práctica diaria, los profesionales deben dedicar una parte considerable de su jornada a tareas administrativas, entre las que destacan la redacción y actualización de historiales clínicos. Esta realidad ha sido reconocida en diversas instituciones y estudios. En el contexto español, según la publicación [iSa22], se estima que entre el 65 y 70 % del tiempo de trabajo de los médicos se destina a labores administrativas, entre ellas la actualización de historiales clínicos. Esta situación no solo impacta negativamente en la calidad laboral de los expertos sanitarios, sino que también reduce el tiempo disponible para la atención a los pacientes, provocando una peor eficiencia en el sistema sanitario.

En los últimos años, avances tecnológicos en Inteligencia Artificial (IA), y en particular, en modelos de procesamiento de lenguaje natural (Natural Language Processing, NLP), han abierto nuevas posibilidades para automatizar tareas complejas y repetitivas dentro del entorno de la salud. En concreto, una de las aplicaciones más prometedoras es la generación automática de informes clínicos, donde sistemas inteligentes basados en IA pueden ayudar a especialistas a redactar textos de manera mucho más fácil, garantizando mayor rapidez, coherencia y precisión en la información.

La posibilidad de utilizar modelos de IA para analizar textos clínicos, resumir su contenido, extraer las características más relevantes o incluso generar informes a partir de datos representa un cambio en la manera en que se gestionan y procesan los historiales clínicos. Esta tecnología no busca sustituir a los profesionales sanitarios, sino ofrecerles una herramienta que les permita ahorrar tiempo.

El propósito de este proyecto es desarrollar una aplicación que facilite la generación de informes clínicos a partir de diagnósticos médicos utilizando IA generativa. Para ello se ha utilizado el conjunto de datos MIMIC-IV [JBS⁺²³], una base de datos clínica ampliamente utilizada en investigación médica. Este conjunto contiene información de una gran cantidad de pacientes ingresados en hospitales, junto con sus diagnósticos escritos por expertos.

1. Introducción

Para desarrollar el proyecto se realizará un proceso de integración de datos, preprocesamiento y limpieza de los mismos. A continuación, se aplicarán técnicas de NLP mediante modelos de generación de lenguaje, que permiten analizar textos clínicos y transformarlos en versiones más concisas y estructuradas. De esta manera, se dará soporte al médico en general para redactar informes más rápidos, consistentes y precisos, reduciendo la carga administrativa y mejorando la eficiencia en los procesos de atención al paciente.

1.2. Problema a resolver

En el contexto de la atención sanitaria, mencionado anteriormente, los profesionales médicos sufren una sobrecarga administrativa. A pesar de los avances en digitalización y la implantación de historiales clínicos electrónicos, la redacción, revisión y actualización de estos sigue siendo una tarea manual, repetitiva y que consume mucho tiempo para estos profesionales. Además, los informes clínicos se redactan en lenguaje libre. Esta variabilidad en la redacción dificulta la homogeneidad y la calidad de los documentos, dificultando su interpretación.

Por otro lado, el volumen de datos clínicos que se genera a diario es cada vez mayor, lo que hace inviable que un profesional revise manualmente todos los registros relacionados con un paciente en un tiempo razonable. Por tanto, existe una necesidad de incorporar herramientas inteligentes que asistan a los sanitarios.

El problema que se aborda en este trabajo se puede resumir en los siguientes puntos:

- La redacción manual de informes clínicos es un proceso costoso en tiempo y repetitivo.
- La variabilidad y falta de estandarización en los textos clínicos dificulta su interpretación.
- La cantidad de información clínica supera la capacidad humana para procesarla de forma eficiente.
- Los sistemas actuales de historiales clínicos electrónicos ofrecen poca automatización de estos.

Ante esta situación, se plantea diseñar un sistema basado en IA que, a partir de textos clínicos existentes, sea capaz de resumirlos y extraer sus características más relevantes.

1.3. Objetivos

1.3.1. Objetivo general del trabajo

El objetivo principal de este TFG es desarrollar una aplicación basada en IA generativa, que utilizando modelos de lenguaje, sea capaz de asistir en la generación de informes clínicos. A diferencia de enfoques centrados en la generación libre de texto, en este proyecto se plantea como objetivo procesar los textos clínicos existentes para resumir, extraer y clasificar su contenido de manera estructurada. Facilitando así la consulta, almacenamiento y reutilización de la información médica.

1.3.2. Objetivos específicos

El objetivo general se ha articulado en torno a la consecución de los siguientes objetivos específicos:

- **OE1.** Estudiar el estado del arte de la IA en salud y de las aplicaciones de generación de texto, especialmente de informes clínicos.
- **OE2.** Explorar y analizar datasets clínicos relevantes, en particular el de MIMIC-IV, así como realizar un proceso de limpieza y preprocesamiento de los datos del dataset.
- **OE3.** Seleccionar e implementar modelos de IA generativa adecuados para la tarea.
- **OE4.** Diseñar y ejecutar pruebas sobre distintas configuraciones de los modelos sobre diagnósticos extraídos del dataset y analizar sus resultados aplicando métricas.

1.4. Planificación y costes del proyecto

1.4.1. Planificación del proyecto

Para abordar el desarrollo de este proyecto, se han dividido los objetivos específicos en las siguientes tareas:

Fase 1: Estudio del estado del arte (OE1)

- Revisión de literatura sobre IA en salud (2 semanas)
- Estudio de modelos de generación de texto (4 semanas)
- Estudio del dataset MIMIC-IV (2 semanas)

Fase 2: Exploración y análisis del dataset (OE2)

- Preprocesamiento y análisis de la demo del dataset (2 semanas)
- Preprocesamiento del dataset (7 semanas)
- Análisis exploratorio de los datos del dataset (4 semanas)

Fase 3: Implementación de modelos de IA generativa (OE3)

- Implementación de los modelos de generación de texto seleccionados (4 semanas)

Fase 4: Ejecución de pruebas y análisis de métricas (OE4)

- Ejecución de las pruebas en los distintos modelos de IA (3 semanas)
- Análisis de los resultados de métricas aplicadas (3 semanas)

Por último, se añade una última tarea, que es la de la redacción de este documento (9 semanas).

Esta planificación ha sido diseñada teniendo en cuenta la complejidad técnica del trabajo, así como el tiempo disponible para su realización. Se muestra un diagrama de Gantt en [Figura 1.1](#) que refleja el desarrollo del proyecto.

1. Introducción

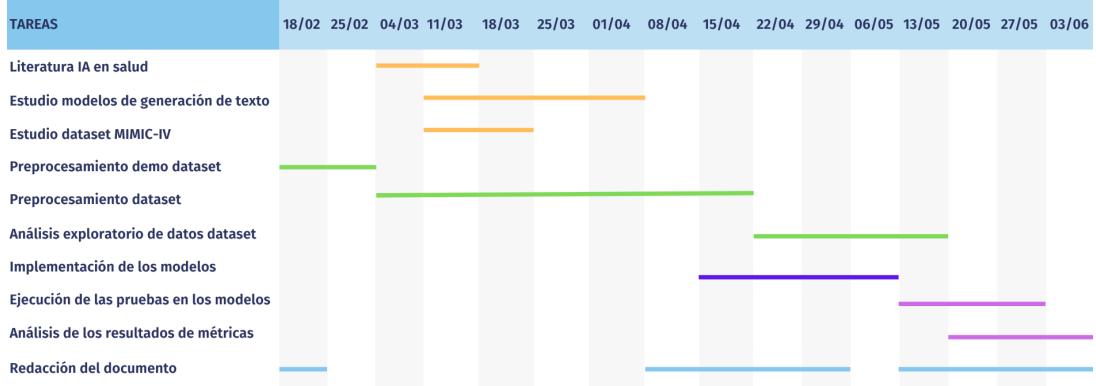


Figura 1.1.: Diagrama de Gantt.

1.4.2. Costes del proyecto

Para calcular los costes del proyecto se ha sumado el coste del personal, teniendo en cuenta a un desarrollador junior y dos tutores; el coste de material y los costes indirectos. Para el coste de material se ha utilizado un ordenador portátil personal y dos servidores proporcionados por la Universidad de Granada.

Coste de personal

- Un desarrollador junior (25€/hora): 20 horas/semana x 4 semanas x 25€/hora = 2.000 €/mes.
- Dos tutores (40€/hora): 2 horas/semana x 4 semanas x 40€/hora x 2 tutores = 640 €/mes.

Coste de material

- Un portatil HP Pavilionn Plus 14: 900€ (precio de compra) / 4 años(vida útil) = 225€/año x 0,5 años = 112,5 € (tiempo de uso exclusivo para el proyecto).
- Licencia de software incluidas en la compra del dispositivo o gratuitas por ser estudiante.
- Servidor con GPU para realizar las pruebas de los modelos: Asus RS720 - 2x Xeon 4210R - 128GB DDR4 2666 - 2xRTX A5000 - 2x1TB - 10TB SATA3: 9.500€ (precio de compra) / 5 años(vida útil) = 1.900€/año x 0,05 años = 95 € (tiempo de uso para las pruebas).
- Servidor para procesar los datos del dataset: Asus RS700-E9-RS4 / 2 x Xeon Silver 4114 10x2,2 Ghz / 128 GB DDR4 2666 / 2x4TB / 10 GBe: 2.800€ (precio de compra) / 6 años(vida útil) = 467€/año x 0,5 años = 233,5 € (tiempo de uso para el proyecto).
- Total: 112,5 + 95 + 233,5 = 441€

Costes indirectos

- Luz: 100€/mes
- Internet: 50€/mes
- Total: 200€/mes × 5 meses = 1.000€

Coste total Coste total = Coste de personal (5 meses × (2.000 + 640€))

$$\begin{aligned} &+ \text{Coste de material (441€)} \\ &+ \text{Costes indirectos (1.000€)} \\ &= 14.641€ \end{aligned}$$

1.5. Estructura de la memoria

1. En el Capítulo 1 se han presentado el contexto, la motivación y los problemas a resolver de este proyecto, así como los principales objetivos que nos hemos propuesto para abordarlos. También, se describe la planificación y costes del proyecto.
2. En el Capítulo 2 se describe el estado del arte en relación a las principales tecnologías, métodos y técnicas utilizadas en salud y medicina, centrándonos sobre todo en modelos de generación de texto. Por último se analiza el uso de MIMIC-IV en la investigación.
3. En el Capítulo 3 se presenta la metodología a desarrollar para el proyecto, dividiéndola en dos grandes fases y justificando el uso de los modelos de lenguaje.
4. El Capítulo 4 se realiza una descripción y análisis profundo del dataset. También se explica como se ha realizado el preprocesamiento y unificación de este dataset, así como el análisis exploratorio de los datos extraídos.
5. En el Capítulo 5 se explica como se han extraído los textos y las distintas métricas utilizadas para analizar cada modelo.
6. El Capítulo 6 describe el diseño y la experimentación de las pruebas realizadas, así como el análisis de estas.
7. En el Capítulo 7 se presentan las conclusiones y trabajos futuros a partir del presente TFG.
8. Finalmente, se presentan las referencias bibliográficas usadas en este trabajo.

2. Estado del arte

El uso de la Inteligencia Artificial ha experimentado un crecimiento significativo en las últimas décadas, especialmente en el ámbito de la salud. En particular, los modelos de IA generativa han demostrado su utilidad en la generación de informes clínicos. Este capítulo revisa el desarrollo de la IA en salud, los modelos de generación de texto y el uso del dataset MIMIC-IV en investigaciones recientes.

En primer lugar, se revisan las principales tecnologías y métodos de IA en salud y medicina. A continuación, se investigan modelos de IA para generación automática de texto y se exploran los diferentes estudios de generación de informes clínicos usando IA generativa. Por último, se estudia sobre el uso del dataset MIMIC-IV en la investigación.

2.1. Inteligencia artificial en salud

Inicialmente, la IA en salud tuvo su comienzo con sistemas expertos como MYCIN [Sho77], centrándose en la toma de decisiones basadas en reglas. Desde entonces, con la aparición del aprendizaje automático y, posteriormente, del aprendizaje profundo, se han desarrollado modelos más complejos capaces de procesar grandes volúmenes de datos clínicos y generar recomendaciones o informes automáticos.

Mediante la automatización de tareas diagnósticas, el análisis de grandes volúmenes de datos clínicos y la mejora en la toma de decisiones, la IA ha demostrado su utilidad en diversas áreas, incluyendo la radiología, la cardiología y la predicción de enfermedades [Top19].

La IA generativa ha comenzado a transformar múltiples áreas de la medicina, particularmente en tareas relacionadas con el procesamiento de lenguaje natural (NLP). Uno de los usos más destacados de la IA generativa es la automatización de informes médicos. Modelos de generación de texto, que serán mencionados en la siguiente sección, han sido entrenados con datos médicos para generar textos clínicamente coherentes. En radiología, modelos encoder-decoder han generado informes sobre imágenes médicas [CSCW20]. Artículos como [VVVUB⁺23] estudian el uso de diferentes modelos de generación de texto utilizando 6 datasets para producir resúmenes de informes clínicos.

2.2. Modelos de IA para generación automática de texto

Los modelos de procesamiento del lenguaje natural han revolucionado la generación de texto en todos los ámbitos, incluyendo la medicina. Entre los modelos más utilizados para la generación de informes clínicos podemos destacar:

- **GPT (Generative Pre-trained Transformer).** Modelos como GPT-3 y GPT-4 han mostrado gran capacidad en la generación de textos coherentes y contextualizados a partir de datos de entrada [RNS⁺18].

2. Estado del arte

- **Llama (Large Language Model Meta AI).** Llama es una familia de modelos de lenguaje optimizados para ser más eficientes en términos de recursos computacionales sin sacrificar calidad. Es de código abierto, lo que facilita su adaptación a datos médicos [TLI⁺[23](#)].
- **Mistral.** Este modelo destaca por su eficiencia y arquitectura innovadora. Consumo poca memoria, ideal para hospitales con limitaciones técnicas y también es de código abierto [JSM⁺[23](#)]. La variante de este modelo, Mixtral [JSR⁺[24](#)], utiliza una arquitectura de tipo Mixture of Experts (MoE), que le permite tener una gran capacidad sin requerir el uso completo del modelo en cada inferencia, lo que lo hace muy eficiente.
- **Qwen.** Desarrollado por Alibaba, Qwen es un modelo multilingüe de código abierto que ha mostrado buenos resultados en comprensión y generación de texto, incluyendo tareas médicas. Su arquitectura eficiente permite su implementación en entornos clínicos con restricciones computacionales [QY⁺[25](#)].
- **Gemma.** Gemma es un modelo ligero y de código abierto desarrollado por Google, enfocado en facilitar la adopción responsable de IA. Su rendimiento en tareas de generación de texto lo hace adecuado para aplicaciones médicas donde la transparencia y la reproducibilidad son fundamentales [TRP⁺[24](#)].
- **Phi.** Desarrollado por Microsoft, Phi es un modelo pequeño y eficiente diseñado para mantener un alto rendimiento en tareas de lenguaje natural, incluso con un número reducido de parámetros. Gracias a su accesibilidad y buen desempeño, se considera una opción prometedora para entornos médicos con recursos limitados [HPBP⁺[24](#)].

Modelo	Número de parámetros	Libre acceso
GPT-3	175B	No
GPT-4	No divulgado	
LLaMA-4	17B	
LLaMA-3	3B, 8B, 70B	Sí
LLaMA-2	7B, 13B, 70B	
Mistral 7B	7B	Sí
Mixtral	12.9B	
Qwen 3	0.6B, 1.7B, 4B, 8B, 14B, 32B	
Qwen 2.5-Omni	7B	Sí
Qwen	72B	
Gemma 2	2B, 9B, 27B	
Gemma 3	1B, 4B, 12B, 27B	Sí
Phi 3	3.8B, 7B, 14B	
Phi 4	3.8B	Sí

Tabla 2.1.: Comparación de modelos de generación de texto según número de parámetros y acceso libre.

Estos modelos se han utilizado en diversas investigaciones para mejorar la generación automática de informes clínicos, optimizando la calidad y reduciendo el tiempo de redacción.

2.3. Generación de informes clínicos usando IA generativa

Diferentes estudios han explorado el uso de IA para la generación de informes clínicos. En áreas como la radiología, donde los modelos han sido entrenados con conjuntos de datos como MIMIC-CXR para producir descripciones automáticas de imágenes médicas [LSX²²].

Por otro lado, varios estudios han explorado la aplicación de modelos de IA a registros electrónicos de salud (EHR). El artículo [BLM²⁴] explora el uso de modelos de lenguaje generativo (LLMs) como ChatGPT para responder a EHR. Dado que la creciente cantidad de mensajes electrónicos está contribuyendo al agotamiento de los clínicos, los autores investigan si los LLMs pueden generar borradores de respuestas para mensajes con tono negativo o emocionalmente cargado.

También, investigaciones recientes han utilizado MIMIC-IV para entrenar modelos capaces de generar informes clínicos basados en datos de pacientes en unidades de cuidados intensivos (UCI), facilitando el seguimiento y la documentación de los casos [JBS²³].

2.4. Uso del dataset MIMIC-IV en la investigación

MIMIC-IV es un conjunto de datos de registros electrónicos de salud del Beth Israel Deaconess Medical Center, que será explicado más a fondo en los siguientes capítulos. Su estructura modular permite aplicaciones avanzadas en predicción de enfermedades y generación de informes.

Estudios clave relacionados con MIMIC-IV que podemos destacar incluyen:

- **Comparison of MIMIC-III and MIMIC-IV for big data analytics of health records [CK23]**. Analiza las mejoras de MIMIC-IV respecto a MIMIC-III, resaltando la introducción de atributos de datos mejorados y un seguimiento más sofisticado de los pacientes en UCI.
- **An Extensive Data Processing Pipeline for MIMIC-IV [GGC²²]**. Presenta una canalización de procesamiento de datos que facilita la extracción, limpieza y preparación de la información contenida en MIMIC-IV.
- **Predicting Prescribed Medications from the MIMIC-IV Event and Measurement Data [Rep23]**. Explora la predicción de medicamentos prescritos a partir de eventos clínicos y mediciones de pacientes. De esta manera, se permite predecir las medicaciones administradas a los pacientes.

Estos trabajos evidencian la importancia de MIMIC-IV como un recurso clave para la investigación en IA aplicada a la salud, especialmente para predicción y uso de IA generativa.

Tras el análisis realizado en el estado del arte, se observa un creciente interés en la aplicación de técnicas de IA en el ámbito sanitario, especialmente en tareas relacionadas con el procesamiento y generación de texto clínico. Existen múltiples trabajos que exploran el uso de modelos de lenguaje para resumir notas médicas o generar fragmentos de informes, especialmente en campos como la radiología o el análisis de EHR.

Asimismo, el dataset MIMIC-IV se ha utilizado en el desarrollo de modelos clínicos gracias a su gran cantidad de datos anonimizados y estructurados. Ha sido empleado para tareas como la predicción de enfermedades, el análisis de medicación o para la generación de informes clínicos automatizados.

2. Estado del arte

No obstante, a pesar de estos avances, se identifican algunas limitaciones que justifican la realización de este proyecto. Aunque existen modelos generativos aplicados al ámbito clínico, la mayoría de estudios se centran en contextos muy específicos. Además, modelos como GPT, Llama, Qwen o Phi han demostrado gran capacidad generativa, pero no se han evaluado ampliamente en entornos clínicos para la generación de informes estructurados.

La aplicación del dataset MIMIC-IV a tareas de generación automática de informes clínicos completos sigue siendo limitada. Muchos artículos utilizan únicamente fragmentos o aspectos numéricos del conjunto de datos, sin explotar completamente el potencial de los textos clínicos.

Por estas razones, se justifica este proyecto de diseñar y desarrollar una aplicación que utilice modelos de IA generativa para crear informes clínicos a partir de diagnósticos. Esta propuesta es coherente porque no hay soluciones que tengan las capacidades que se van a desarrollar.

3. Metodología

El desarrollo de este proyecto se ha estructurado principalmente en dos grandes fases: la unificación y preprocesamiento de los datos, y el diseño de un sistema de generación de informes clínicos utilizando IA generativa. A continuación, se describen de manera general estas fases, que más adelante serán desarrolladas en detalle.

3.1. Unificación y preprocesamiento de datos clínicos

Para llevar a cabo la generación de informes, necesitamos un conjunto de datos clínicos de base. En nuestro caso, se ha utilizado el dataset MIMIC-IV, un conjunto de datos que contiene información sobre pacientes ingresados en un hospital, así como sus historiales clínicos.

El primer paso ha consistido en la selección, extracción y unificación de las distintas tablas del conjunto de datos. Estas tablas contienen datos generales del paciente, medicamentos suministrados, áreas visitadas y enfermedades sufridas, entre otras. Este proceso ha implicado:

- Realizar un análisis profundo de las tablas y sus relaciones en el dataset.
- Limpiar datos inconsistentes o incompletos.
- Unificación de todos los datos seleccionados.
- Realizar un análisis exploratorio de los datos preprocesados.

3.2. Generación automática de informes clínicos

Esta segunda fase se centra en el uso de modelos de IA para generar informes clínicos a partir de un diagnóstico dado. El enfoque adoptado consiste en resumir el contenido clínico para extraer las secciones más relevantes y clasificar la información según su categoría. Para ello, se ha trabajado con modelos de lenguaje natural (NLP) basados en arquitecturas transformer. Las principales tareas llevadas a cabo en esta etapa son:

- Selección y preparación de los informes clínicos.
- Generación de prompts para la entrada al modelo de lenguaje.
- Generación del informe clínico a partir de los modelos.
- Evaluación de resultados mediante extracción de métricas que evalúan el desempeño de los modelos a la hora de generar las respuestas.

En la [Figura 3.1](#) se muestra el flujo de trabajo adoptado, en la que se muestran los pasos de la primera fase de color morado y los de la segunda de color verde.

3. Metodología

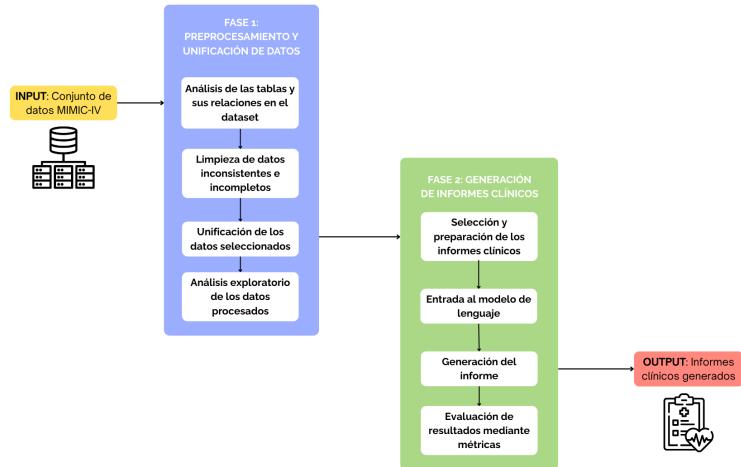


Figura 3.1.: Flujo de trabajo del proyecto.

3.3. Justificación de los modelos de lenguaje utilizados

Para abordar la generación automática de los informes clínicos, se hace uso de modelos de lenguaje basados en arquitecturas transformer. Estos modelos son ampliamente utilizados en tareas de procesamiento de lenguaje natural por su alta capacidad de comprensión semántica, generación de texto y adaptabilidad a distintos dominios.

Entre todas las alternativas disponibles, se han considerado modelos preentrenados y adaptables a cualquier tarea, en especial los denominados Small Language Models (SLM). A diferencia de los Large Language Models (LLM), estos modelos procesan y generan lenguaje humano con tamaños pequeños y menos parámetros que los LLM. La elección de los modelos ha seguido los siguientes criterios:

- Disponibilidad y soporte:** Se ha optado por modelos accesibles desde bibliotecas como Hugging Face, con documentación suficiente, licencias abiertas y facilidad de implementación.
- Rendimiento computacional:** Se han tenido en cuenta modelos con pocos parámetros para poder realizar las pruebas en un tiempo considerable.
- Actualidad:** Se ha optado por los modelos más nuevos de diferentes desarrolladores.
- Adaptabilidad:** Se han tenido en cuenta la versatilidad de los modelos para ser aplicados a distintas tareas y a distintos dominios.

En el caso particular de este proyecto, como los textos extraídos del dataset están escritos en inglés, se ha optado por modelos entrenados originalmente en este idioma, garantizando la mejor comprensión del lenguaje clínico en estos documentos.

Por tanto, la elección final se basa en la actualidad, disponibilidad, adaptabilidad y rendimiento computacional de los modelos.

4. Conjunto de datos, descripción y unificación

En este capítulo se explica con detalle la primera fase que se ha realizado, mencionada en el capítulo anterior. En la [Figura 4.1](#) se muestra el flujo de trabajo adaptado para esta fase. Para ello, se hace un análisis de las tablas y se describe el dataset utilizado, MIMIC-IV. Más tarde, limpian los datos y se unifican, explicado en la sección de preprocesamiento y unificación de los datos del dataset. Por último, se realiza un análisis exploratorio de los datos.

4.1. Análisis y descripción del dataset

Comenzamos trabajando con la demo del dataset MIMIC-IV Clinical Database. Esta demo está compuesta por registros médicos electrónicos anónimos de pacientes ingresados en el Centro Médico Beth Israel Deaconess. Contiene un subconjunto de 100 pacientes, lo cual resulta ideal para comenzar a trabajar con unos datos más recogidos, por temas de almacenamiento y computación. El conjunto de datos incluye contenido similar a MIMIC-IV, pero excluye notas clínicas de texto libre. A cada persona en MIMIC-IV se le asigna un valor único `subject_id`.

Más tarde se comenzó a trabajar en el dataset grande. MIMIC-IV es una base de datos relacional que contiene hospitalizaciones reales de pacientes ingresados en un centro médico académico de tercer nivel en Boston, Massachusetts, EE. UU. MIMIC-IV contiene información completa de cada paciente durante su estancia en el hospital: análisis de laboratorio, medición administrada, constancias vitales registradas, etc. Este se compone de tres principales módulos: **hosp**, **icu** y **note**:

- El módulo **hosp** proporciona todos los datos adquiridos del historial médico electrónico del hospital. Estas mediciones se registran predominantemente durante la estadía en el hospital, aunque algunas tablas también incluyen datos de fuera del hospital (por ejemplo, pruebas de laboratorio ambulatorias en `labevents`). La información incluye detalles del paciente y de la admisión (`patients`, `admissions`, `transfers`), mediciones de laboratorio (`labevents`, `d_labitems`), cultivos microbiológicos (`microbiologyevents`), órdenes del proveedor (`poe`, `poe_detail`), administración de medicamentos (`emar`, `emar_detail`), prescripción de medicamentos (`prescripciones`, `farmacia`), información de facturación del hospital (`diagnoses_icd`, `d_icd_diagnoses`, `procedures_icd`, `d_icd_procedures`, `hcpcsevents`, `d_hcpcs`, `drgcodes`) e información relacionada con el servicio del hospital (`services`). Todos los nombres mencionados entre paréntesis corresponden a los archivos contenidos en este módulo. Estos archivos son de extensión `.csv.gz`, en los que se guardan distintos tipos de atributos según la tabla recogida. La mayoría contienen el atributo `subject_id`, seguido de las demás columnas separadas por comas, como por ejemplo, `hadm_id`, `emar_id`, `charttime`, entre otros. Los tipos de datos que han sido utilizados son: enteros, flotantes, cadenas de caracteres y formato fecha-hora.
- El módulo **icu** contiene información recopilada del sistema de información clínica que se utiliza en la UCI. Contiene datos provenientes del sistema de información clínica del BIDMC: MetaVision (iMDSoft). Las tablas de MetaVision se desnormalizaron para crear

4. Conjunto de datos, descripción y unificación

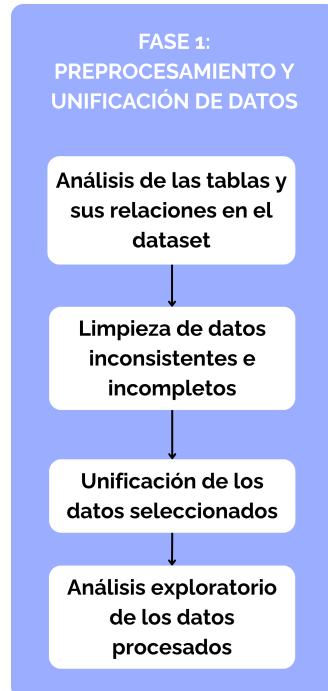


Figura 4.1.: Diagrama de flujo de trabajo del preprocesamiento y unificación del dataset.

un esquema en estrella donde las tablas icustays y d_items se vinculan a un conjunto de tablas de datos, todas con el sufijo “events”. Los datos documentados en el módulo de UCI incluyen entradas intravenosas y de fluidos (inpuutevents), ingredientes de las entradas mencionadas anteriormente (ingredienteevents), salidas de pacientes (outputevents), procedimientos (procedureevents), información documentada como fecha u hora (datetimemevents) y otra información registrada (chartevents). Todas las tablas de eventos contienen una columna stay_id que permite la identificación del paciente de UCI asociado en icustays, y una columna itemid que permite la identificación del concepto documentado en d_items. Asimismo, como en el módulo hosp, estos archivos son de extensión .csv.gz , en los que se guardan distintos tipos de atributos según la tabla recogida. También contienen el atributo subject_id, seguido de las demás columnas separadas por comas, como por ejemplo, hadm_id, emar_id, item_id, amount, entre otros. Los tipos de datos que han sido utilizados son: enteros, flotantes, cadenas de caracteres y formato fecha-hora.

- El módulo **note** contiene notas clínicas de texto libre desidentificadas para pacientes hospitalizados. Los archivos contenidos en este módulo son: discharge (resúmenes de alta por hospitalizaciones), discharge_detail (información auxiliar para resúmenes de alta), radiology (informes de radiología) y radiology_detail (información auxiliar para notas de radiología). Estas notas clínicas están escritas en el idioma inglés, de manera que trabajaremos con estos sin traducirlos.

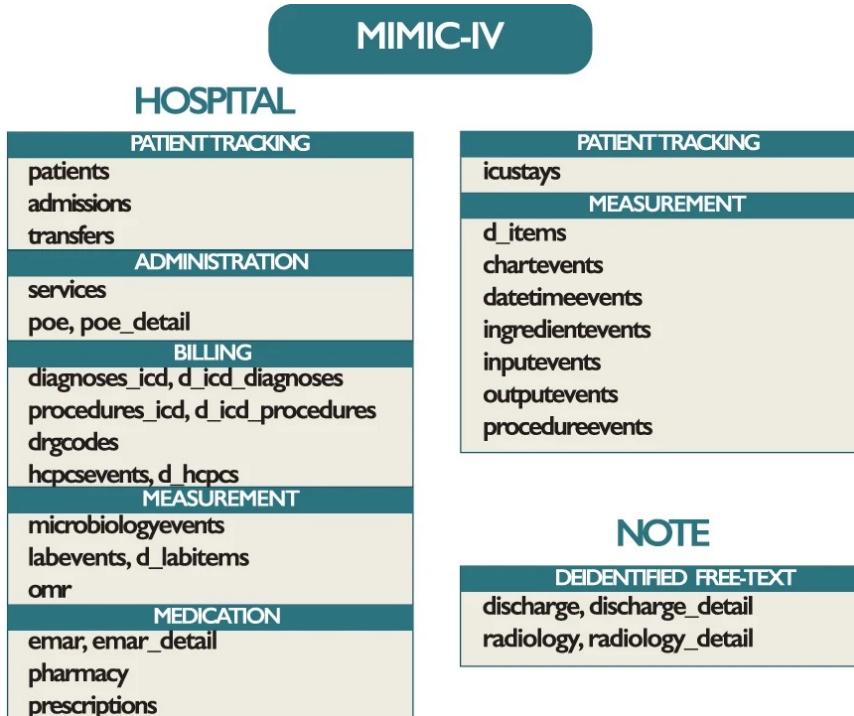


Figura 4.2.: División de los módulos y tablas del dataset. Fuente: [JBS⁺23].

4.1.1. Anonimización de datos

MIMIC-IV es una base de datos que contiene datos de pacientes ingresados entre 2008 y 2022. Dado el carácter sensible de la información médica, este dataset ha sido cuidadosamente anonimizado para proteger la privacidad de los pacientes, cumpliendo con la normativa legal aplicable en Estados Unidos, denominada HIPAA (Health Insurance Portability and Accountability Act). Según se define en [HIP]: "HIPAA es una ley federal que establece requisitos de privacidad y seguridad de los datos para las organizaciones que deben salvaguardar la información médica protegida de otras personas". Esta ley protege la información de salud personal de los pacientes y regula cómo los proveedores de atención médica, los planes de salud y otras entidades pueden usar y divulgar esta información.

La anonimización se realiza mediante varios procedimientos:

- **Eliminación de identificadores directos:** Todas las variables dependientes de una persona, como nombres, número de la Seguridad Social, direcciones, números de teléfono y direcciones de correo electrónico han sido eliminados del conjunto de datos.
- **Desidentificación de fechas y edad:** Las fechas relacionadas con admisiones, altas y otros eventos clínicos han sido desplazadas temporalmente. Para desidentificar adecuadamente la fecha exacta de las estancias de los pacientes, la tabla patients contiene una

4. Conjunto de datos, descripción y unificación

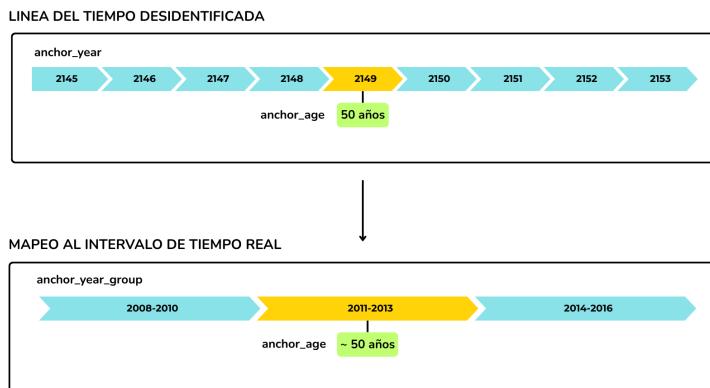


Figura 4.3.: Diagrama de la relación entre la cronología desidentificada y la real.

columna `anchor_year`, un año desplazado. Esta columna ancla los datos almacenados a un año dentro de su cronología desidentificada (p. ej. 2150). En este año desidentificado, la edad del paciente se proporciona en la columna `anchor_age` y el año real aproximado de ingreso se proporciona en la columna `anchor_year_group`. Por ejemplo, si el `anchor_age` de un paciente es 50 y su `anchor_year` es 2150, entonces tenía 50 años en el año 2150. Si el `anchor_year_group` de este paciente es 2011-2013, entonces sabemos que cualquier hospitalización ocurrida en el año desidentificado 2150 en realidad ocurrió en algún momento entre 2011 y 2013, y que tenía 50 años durante este periodo. Un ejemplo visual se encuentra en [Figura 4.3](#). Además, los pacientes mayores de 89 años son agrupados en una categoría de "90 o más", etiquetados con edad de 91 años. Esto se verá reflejado más adelante en el análisis exploratorio de los datos. Finalmente, la fecha de fallecimiento de un paciente está disponible en la columna `dod`. Las fechas de fallecimiento se censuran al año del último alta hospitalaria del paciente.

- **Codificación de identificadores internos:** Se utilizan códigos numéricos generados aleatoriamente para identificar pacientes, episodios y localizaciones, sin posibilidad de rastreo directo hacia las identidades reales. Un ejemplo de esta codificación es el atributo `subject_id`.
- **Desidentificación de las notas:** Como parte del proceso de desidentificación, se han eliminado las secciones de Historial Social e Instrucciones de Alta. Estas secciones solían contener información social y logística irrelevante para la atención médica, pero que presentaba un mayor riesgo de reidentificación en comparación con otras secciones.

Por último, cabe destacar que para acceder al dataset es necesario firmar un acuerdo de uso de datos (DUA). En él se exige a los usuarios proteger adecuadamente el conjunto de datos, no intentar reidentificar a las personas, no compartir los datos y reportar cualquier problema relacionado con la desidentificación.

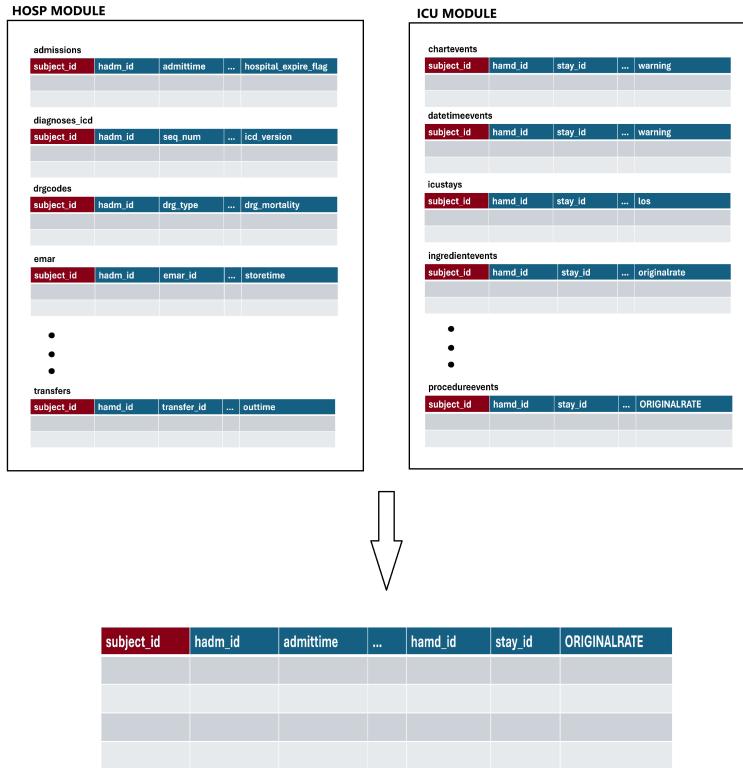


Figura 4.4.: Relación de los datos del dataset para la unificación en el preprocesamiento.

4.2. Preprocesamiento y unificación

Para comenzar la manipulación de los datos, hemos partido de la demo del dataset (mimic-iv-clinical-database). A partir del zip proporcionado, hemos cargado las distintas tablas en formato csv dentro de los dos módulos mencionados anteriormente. Hemos recorrido todos los archivos que contienen la columna `subject_id`, que será nuestra clave principal para distinguir a los distintos pacientes. Estas tablas provenientes de hosp y de icu han sido concatenadas en un nuevo dataframe.

Todas estas ejecuciones se han realizado con la librería Pandas, una librería de Python especializada en la manipulación y el análisis de datos. Ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales.

Una vez realizadas las pruebas con la demo de este dataset, nos pasamos al conjunto grande, con un tamaño de 9,82 GB comprimido. Inicié una prueba para ver si podía preprocesarlo en mi ordenador personal, pero este tardaba mucho tiempo y no llegaba ni siquiera a descomprimir el primer archivo. También intenté ejecutar código en Google Colab pero no tuve éxito. Utilicé otras librerías para preprocesamiento de grandes archivos, como Dask, en la que la descompresión de los archivos se hacía en cuestión de segundos, pero a la hora de concatenarlos seguía teniendo bastantes problemas.

4. Conjunto de datos, descripción y unificación

Como necesitaba un dispositivo con mejores prestaciones, se me concedió el uso de un servidor de la universidad. Al iniciar la descompresión y concatenación de los distintos archivos, tuve varios problemas debido al gran tamaño del dataset. Primero, al ejecutar el código utilizado con la demo, el kernel fallaba a la media hora de ejecutar la mayor parte de los archivos, debido a varios problemas de RAM en el servidor.

Tras varios días de investigación, comencé a utilizar librerías que permitieran las paginación de datos, entre las que usé Dask. Esta es una biblioteca de Python de código abierto para computación paralela. Dask escala código Python desde máquinas locales multinúcleo a grandes clústeres distribuidos en la nube. Al comenzar a utilizarla, funcionaba muy bien cargando archivos, ya que tardaba muy poco tiempo y las líneas de código que había que cambiar con respecto al preprocesamiento de la demo eran escasas, pero el problema surgió otra vez con la concatenación de los distintos archivos.

Para solucionar este problema, volví otra vez a utilizar la biblioteca Pandas, en la que esta vez leería los archivos en distintos chunks y los guardaría en varios subconjuntos.

4.2.1. Preprocesamiento módulo hosp

Para preprocesar los archivos del módulo hosp, estos se cargaron y fueron divididos en 4 grandes subconjuntos. En la tabla [Tabla 4.1](#) podemos ver el número de filas y columnas que han sido guardadas y concatenadas de este módulo. Para ejecutar las celdas de código que extraen estas tablas se tardó alrededor de 3 horas y media en total.

Nombre del subconjunto	Número de filas	Número de columnas
hosp_dataset_sub1	88,762,729	34
hosp_dataset_sub2	112,176,190	63
hosp_dataset_sub3	205,933,437	48
hosp_dataset_sub4	4,369,965	32

Tabla 4.1.: Número de filas y columnas de los subconjuntos extraídos del módulo hosp.

4.2.2. Preprocesamiento módulo icu

Para preprocesar los archivos del módulo icu, estos se cargaron y fueron divididos en 4 grandes subconjuntos como en hosp. Como el archivo chartevents.csv ocupaba tanto espacio, fue dividido en varias partes. En la tabla [Tabla 4.2](#) podemos ver el número de filas y columnas que han sido guardadas y concatenadas de este módulo. Para ejecutar las celdas de código que extraen estas tablas se tardó alrededor de 3 horas en total.

Nombre del subconjunto	Número de filas	Número de columnas
icu_dataset_sub1	118,356,886	11
icu_dataset_sub2	111,762,419	33
icu_dataset_sub3	94,347,938	27
icu_dataset_sub4	149,979,761	11

Tabla 4.2.: Número de filas y columnas de los subconjuntos extraídos del módulo icu.

4.2.3. Preprocesamiento módulo notes

Para preprocesar los archivos del módulo notes, estos se cargaron y fueron concatenados directamente, ya que se pudo realizar en poco tiempo. Este módulo tiene 30,085,407 filas y 11 columnas y tardó unos 40 minutos en ejecutarse. Más tarde, se unificó con el archivo full_diagnostic_with_icd10, que contiene 318,310 filas y 22 columnas. Este archivo contiene también la columna subject_id como todos los archivos del dataset, y además, tiene los nombres de las enfermedades que son descritas en notes divididas en varias categorías, sirviéndonos para que el modelo aprenda más sobre la enfermedad. Finalmente, este módulo contiene 9,203,717 filas y 31 columnas, ocupando un total de 6.61 GB.

4.2.4. Filtración subject_id presentes en módulo notes

En el módulo notes no se encuentran recogidos los datos de todos los pacientes presentes en los módulos hosp e icu, es decir, hay subject_id recogidos en hosp e icu que no lo están en notes. Por tanto, como estos datos no van a ser utilizados en el modelo, van a ser filtrados. Para ello, se extraen los valores únicos de subject_id en un fichero denominado subject_ids.txt, conteniendo los 248,649 valores únicos presentes en notes.

Para filtrar este atributo, se ha cargado el archivo subject_ids.txt y cada subconjunto de los módulos hosp e icu. A partir de ahí, recorriendo cada fila del subconjunto se ha comprobado si el atributo subject_id de la fila existe en el archivo. Si está presente, se queda en el subconjunto, sino es eliminada toda la fila. Finalmente, los subconjuntos son reducidos, almacenando el número de filas recogidas en la tabla [Tabla 4.3](#) y el gráfico [Figura 4.5](#), demostrando que aproximadamente el 18 % del dataset ha sido eliminado, de 885,689,325 filas a 726,948,615.

Nombre del subconjunto	Número de filas (antes)	Número de filas (después)
hosp_dataset_sub1	88,762,729	74,990,340
hosp_dataset_sub2	112,176,190	89,822,760
hosp_dataset_sub3	205,933,437	171,703,463
hosp_dataset_sub4	4,369,965	3,634,667
icu_dataset_sub1	118,356,886	94,840,246
icu_dataset_sub2	111,762,419	89,948,910
icu_dataset_sub3	94,347,938	77,276,054
icu_dataset_sub4	149,979,761	124,732,175

Tabla 4.3.: Número de filas anteriores y posteriores a la filtración del atributo subject_id.

4.2.5. Concatenación de todos los subconjuntos

Finalmente, todos los subconjuntos son concatenados, añadiendo todas las filas de los 4 subconjuntos de hosp, los 4 de icu y el de notes. Son guardados en el archivo denominado final_dataset.csv, ocupando un total de 98 GB. El tiempo que ha tardado en añadirse las 736,152,332 filas ha sido recogido en la tabla [Tabla 4.4](#). Finalmente, todos los pasos realizados para el preprocesamiento quedan recogidos en [Figura 4.6](#).

4. Conjunto de datos, descripción y unificación

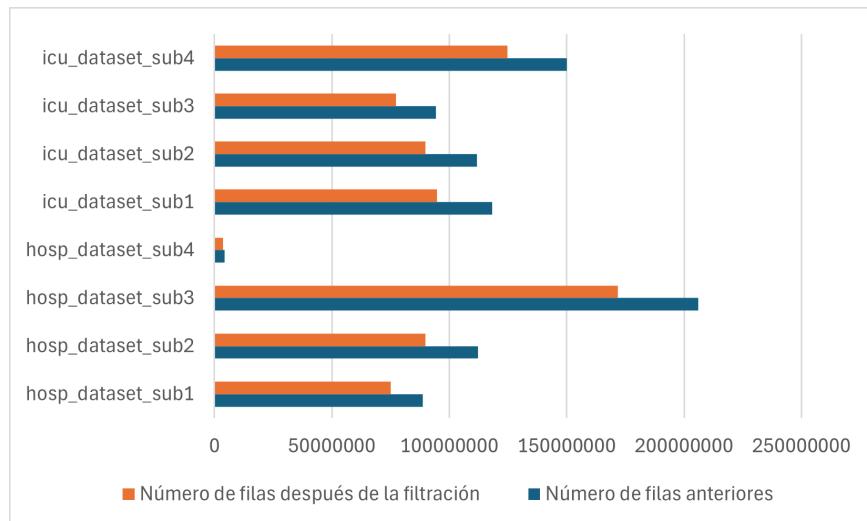


Figura 4.5.: Gráfico de comparación de filas contenidas en los subconjuntos antes y después de la filtración.

Nombre del subconjunto	Tiempo (minutos)	Número de filas
hosp_dataset_sub1	23	74,990,340
hosp_dataset_sub2	43	89,822,760
hosp_dataset_sub3	82	171,703,463
hosp_dataset_sub4	1	3,634,667
icu_dataset_sub1	12	94,840,246
icu_dataset_sub2	25	89,948,910
icu_dataset_sub3	19	77,276,054
icu_dataset_sub4	14	77,276,054
dataset_notes_unificado	6	124,732,175

Tabla 4.4.: Comparación del tiempo y filas que se ha tardado en concatenar de cada subconjunto en el conjunto final.

4.3. Análisis exploratorio de datos

Antes de abordar la generación de informes clínicos, se ha realizado un análisis exploratorio del dataset con el fin de comprender mejor la estructura, distribución y características principales de las variables más relevantes.

Debido a la gran cantidad de datos recogidos en el dataset final, para realizar el análisis se han utilizado 2 tablas del módulo hosp y el módulo notes completo. Para visualizar y analizar los datos para nuestro caso, creemos que la mejor manera para no consumir mucho tiempo y espacio es extraer las tablas patients y admission de hosp, que contienen datos relevantes de los pacientes y sus enfermedades, así como el módulo notes, que contiene los diagnósticos de los pacientes.

Para unificar estos 3 subconjuntos, se ha realizado un merge de ellos con las claves primarias subject_id y hadm_id. Finalmente, este conjunto está formado por 546,028 filas y 47

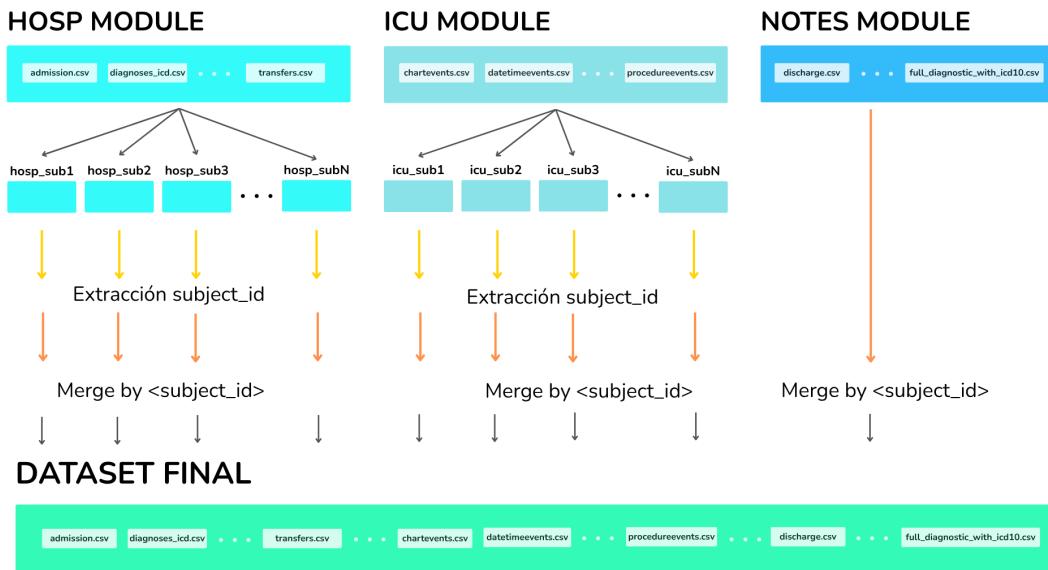


Figura 4.6.: Pasos realizados durante el preprocesamiento para la división y concatenación de los módulos del dataset.

columnas.

4.3.1. Distribución de la edad de los pacientes

Se ha construido un histograma con las edades de los pacientes, utilizando la columna anchor_age. En la Figura 4.7 podemos observar un claro predominio de pacientes adultos mayores, principalmente de edades entre 56 y 66 años.

Además, vemos un gran salto de 89 a 91 años. Esto ha sido comentado anteriormente en la anonimización de datos, los pacientes con edades mayores a 89 años han sido clasificados con edad de 91 años. De hecho, no existe ni la edad 90 ni ninguna mayor a 91 en la gráfica porque no hay ningún registro con esas edades.

4.3.2. Ubicación de alta del paciente

Otra variable de interés para explorar es la ubicación de alta del paciente, recogido en la columna discharge_location, que indica donde fue trasladado el paciente tras el alta. Este análisis permite tener una idea del estado clínico final de los pacientes. En el gráfico Figura 4.8 se aprecia que la categoría más frecuente es HOME, indicando que casi 200,000 pacientes fueron enviados a casa. También 100,000 pacientes fueron dados de alta a su casa con cuidados, y casi 12,000 fallecieron.

4.3.3. Distribución por género

Otra variable representada es el género, recogido en la variable gender, con dos categorías principales: "M"(masculino) y "F"(femenino). En la Figura 4.9, la proporción observada está

4. Conjunto de datos, descripción y unificación

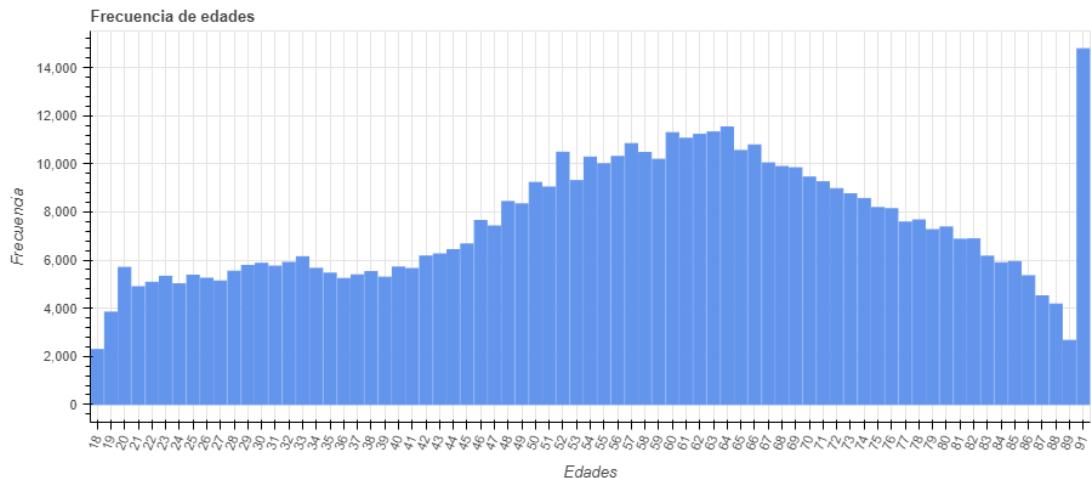


Figura 4.7.: Distribución de edades de los pacientes.

razonablemente equilibrada, aunque con un leve predominio de pacientes femeninas.

También se ha estudiado la relación de la variable edad con la variable género. Para ello se ha dividido la edad en bins de 10 en 10 años y se ha agrupado por F(femenino) y M(masculino). En la imagen [Figura 4.10](#) se observa un claro predominio de mujeres que acuden al hospital entre las edades de 10 a 30 años. A partir de los 30 hasta los 80 superan los hombres en visitas, mientras que a partir de los 80 vuelven a predominar las mujeres.

4.3.4. Enfermedades más frecuentes

Una variable muy importante para poder clasificar y entender de qué tratan los textos es la enfermedad que ha sufrido un paciente. Para ello se ha realizado una clasificación de las 10 enfermedades más frecuentes entre los pacientes, agrupadas por género. En la [Figura 4.11](#) se observan estas 10 enfermedades, que en la mayoría son contraídas por hombres con bastante diferencia que mujeres. Entre las 3 más frecuentes encontramos:

- **Acute kidney failure:** Insuficiencia renal aguda.
- **Artherosclerotic heart disease of native coronary artery:** Enfermedad cardíaca arteriosclerótica de la arteria coronaria nativa.
- **Cellulitis and acute lymphangitis of other parts of limb:** Celulitis y linfangitis aguda de otras partes de la extremidad.

En la [Figura 4.12](#) se pueden visualizar las 10 enfermedades más frecuentes entre pacientes fallecidos y la cantidad de estos que la han sufrido. Entre las 3 más frecuentes encontramos:

- **Sepsis, unspecified organism:** Sepsis, organismo no especificado.
- **Encounter for antineoplastic chemotherapy and immunotherapy:** Encuentro para quimioterapia antineoplásica e inmunoterapia.
- **Acute kidney failure:** Insuficiencia renal aguda.

4.3.5. Análisis individual de un paciente

Como complemento al análisis general, se ha seleccionado un paciente individual para hacer una revisión detallada de su historia clínica. El paciente es identificado por el subject_id 10012853. En la tabla de la Figura 4.13 se observa que este paciente fue admitido 6 veces en el hospital en un periodo de 4 años. Su admisión inicial fue el 5 de abril del año desidentificado 2175 y duró 5 días. El lugar de admisión fue a través de la sala de emergencia y el lugar de alta fue a casa. En la imagen Figura 4.14 se visualiza una línea del tiempo que recoge las 6 visitas del paciente, junto con la duración de cada una.

4.3.6. Diagnóstico de un paciente

Como se ha visto, este conjunto de datos contiene miles de textos generados por especialistas médicos. En el apéndice Sección A.1 se muestra un ejemplo de informe clínico real correspondiente a un paciente con una enfermedad del sistema respiratorio.

4.3.7. Wordcloud

Una vez mostrado el diagnóstico, se ha explorado el contenido más frecuente de los informes clínicos, generando nubes de palabras (wordclouds). En concreto, se han seleccionado informes relacionados con enfermedades del sistema circulatorio y del sistema respiratorio.

Analizando los wordclouds presentados en Figura 4.15 y Figura 4.16, podemos visualizar que ambos muestran en gran medida palabras como mg, PO, DAILY o Tablet. Esto refleja la repetición de la frecuencia de la medicación y dosificación de los distintos fármacos escritos en los diagnósticos. Además, también se ve en las dos nubes la palabra patients ya que se referencia mucho al paciente en cuestión para describir su estado, medicación, molestia o cualquier característica, para no revelar su identidad.

El wordcloud del sistema circulatorio destaca términos específicos del campo como aortic valve, coronary artery, left ventricular o mitral valve. Por lo que destaca terminología relacionada con válvulas y anatomía cardíaca. En el wordcloud del sistema respiratorio vemos términos como abnormal findings, pleural effusion o chest pain, enfocándose en síntomas respiratorios y hallazgos anómalos.

En resumen, aunque los wordclouds revelan que cada sistema tiene su terminología específica, comparte un marco común de manejo clínico, medicación, farmacología y monitoreo de parámetros. Esto apoya nuestra metodología genérica respecto al tratamiento de los textos, independientemente de la especialidad.

4. Conjunto de datos, descripción y unificación

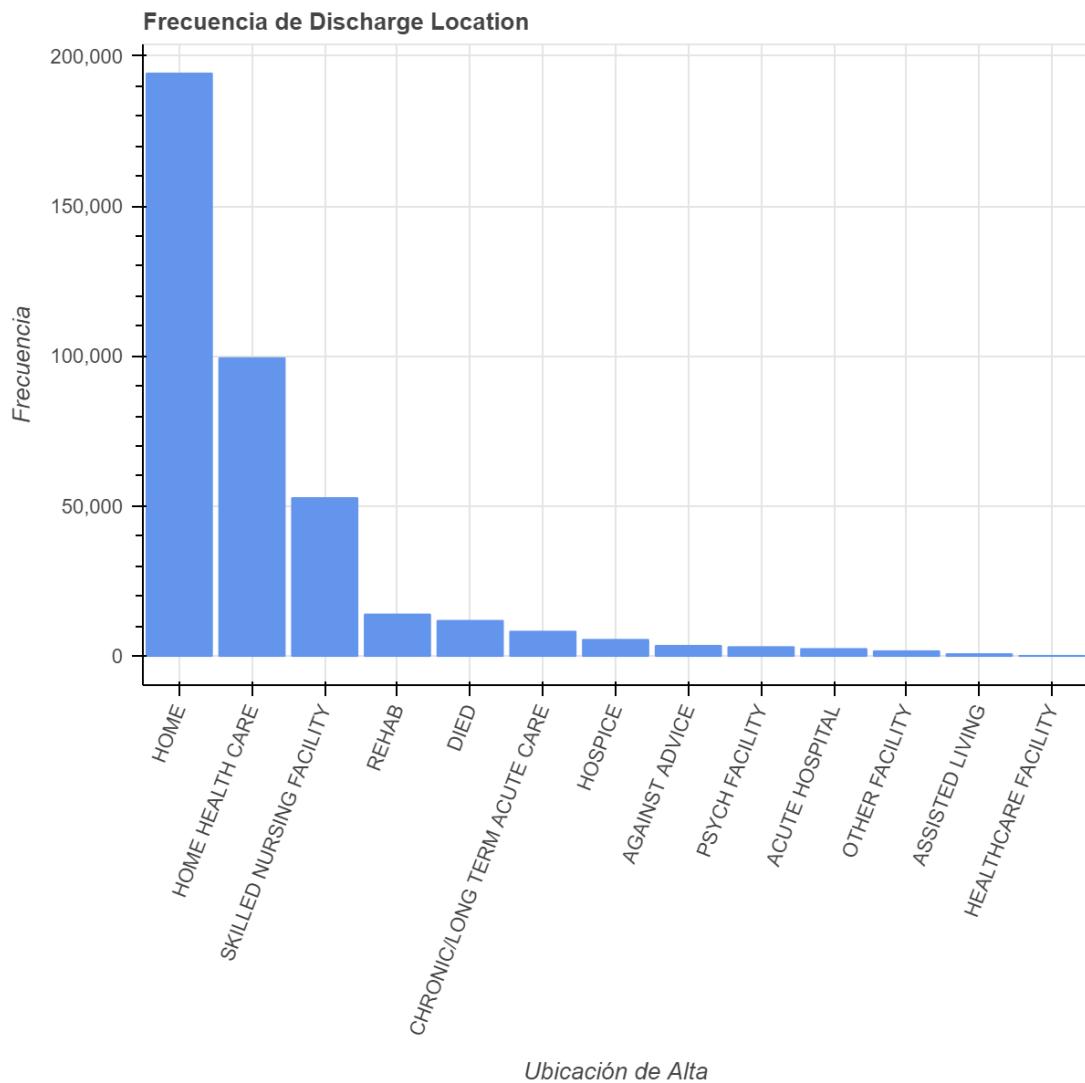


Figura 4.8.: Distribución de la ubicación de alta de los pacientes.

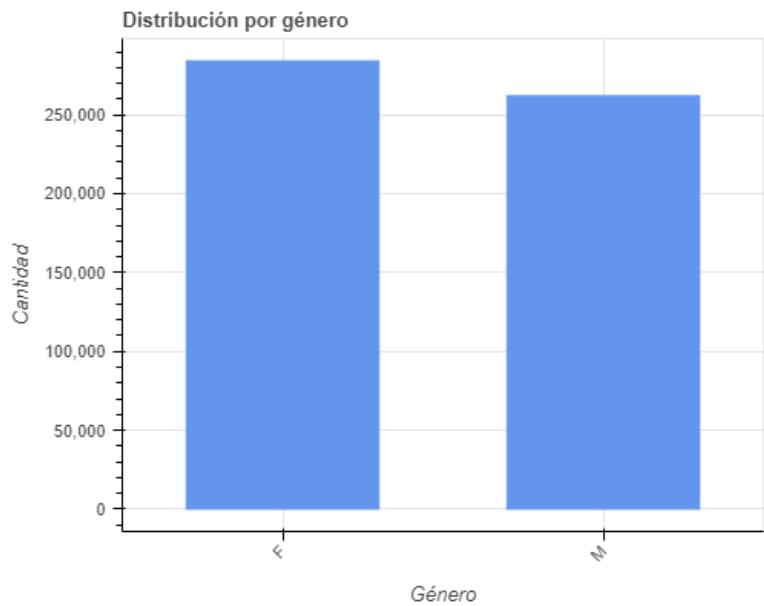


Figura 4.9.: Distribución del género de los pacientes.

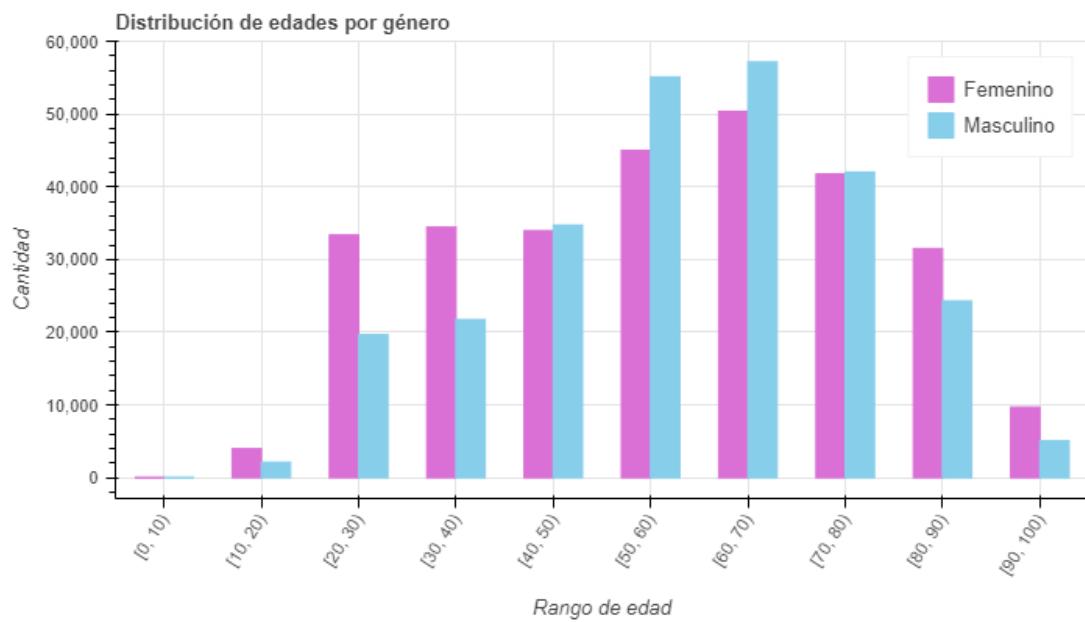


Figura 4.10.: Distribución de las edad agrupada por género de los pacientes.

4. Conjunto de datos, descripción y unificación

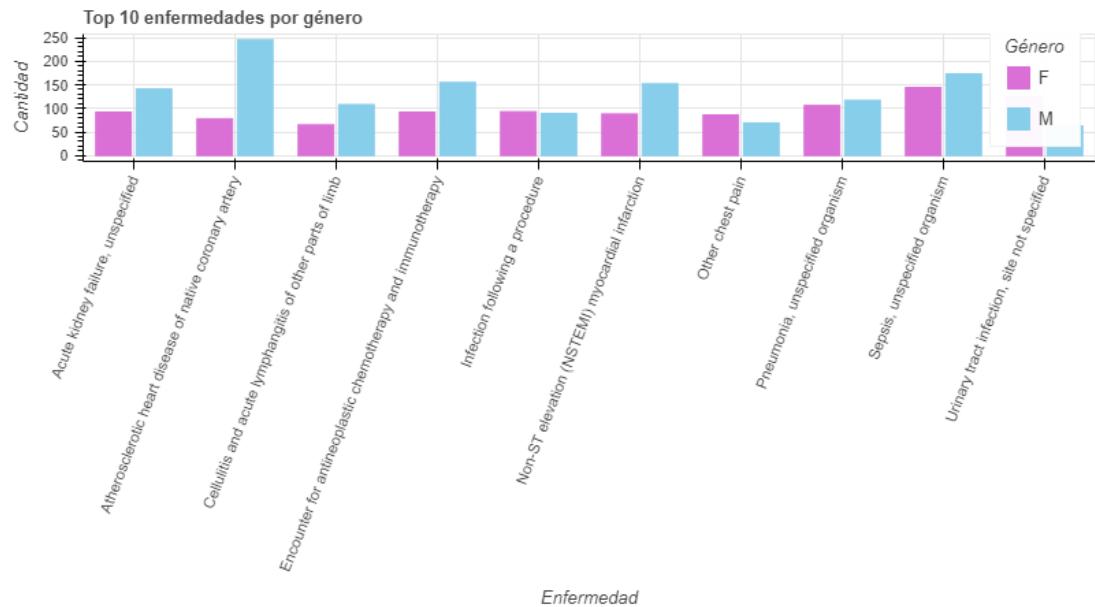


Figura 4.11.: Top 10 enfermedades más frecuentes agrupadas por genero.

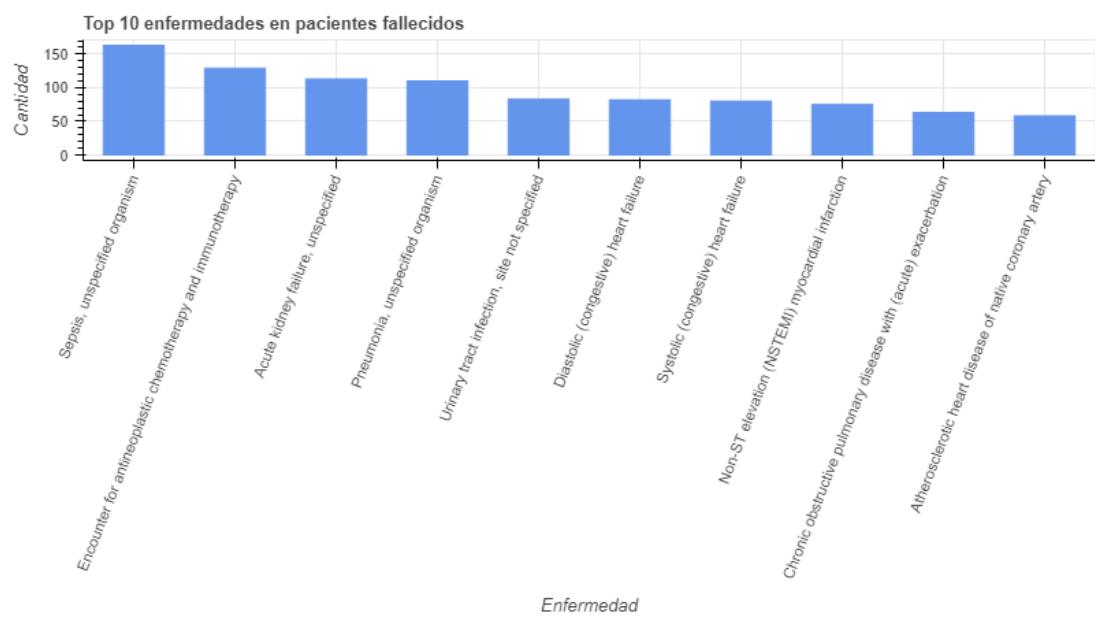


Figura 4.12.: Top 10 enfermedades más frecuentes entre los pacientes fallecidos.

4.3. Análisis exploratorio de datos

	subject_id	hadm_id	admittime	dischtime	admission_location	discharge_location
724	10012853	26369609	2175-04-05 15:36:00	2175-04-10 16:55:00	EMERGENCY ROOM	HOME
722	10012853	22539296	2176-06-06 18:09:00	2176-06-08 18:30:00	EMERGENCY ROOM	HOME
723	10012853	22896692	2176-08-11 15:17:00	2176-08-11 17:35:00	EMERGENCY ROOM	nan
726	10012853	27882036	2176-11-25 21:28:00	2176-12-03 15:24:00	EMERGENCY ROOM	SKILLED NURSING FACILITY
725	10012853	26739864	2177-02-11 07:15:00	2177-02-12 16:20:00	PHYSICIAN REFERRAL	nan
721	10012853	28457729	2177-11-03 09:30:00	2177-11-04 15:06:00	PHYSICIAN REFERRAL	nan

Figura 4.13.: Tabla de las admisiones en el hospital para el paciente 10012853.

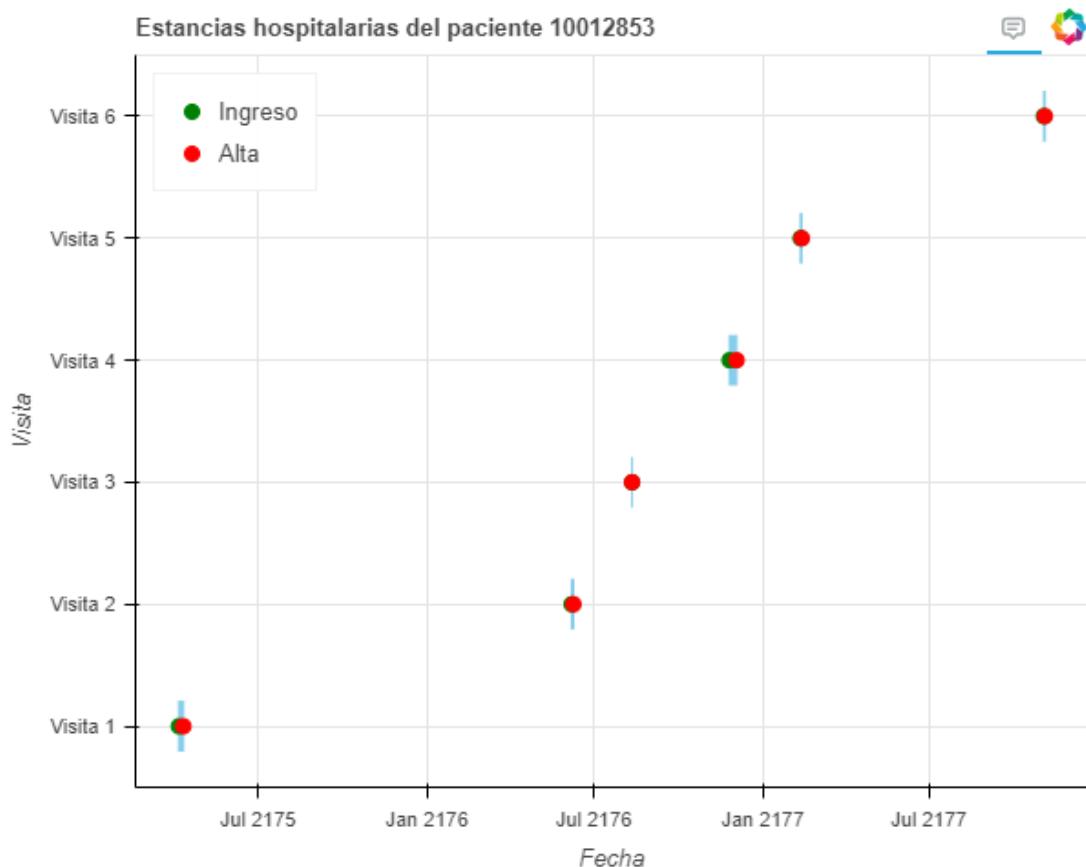


Figura 4.14.: Visitas hospitalarias del paciente 10012853.

4. Conjunto de datos, descripción y unificación



Figura 4.15.: Nube de palabras generada a partir de informes clínicos relacionados con enfermedades del sistema circulatorio.



Figura 4.16.: Nube de palabras generada a partir de informes clínicos relacionados con enfermedades del sistema respiratorio.

5. Generación automática de informes clínicos

En este capítulo se explica brevemente el proceso llevado a cabo para la generación automática de los informes clínicos, así como las distintas métricas utilizadas para evaluar el desempeño de los modelos para generar estos informes.

5.1. Procedimiento

Para realizar el proceso de generación automática de informes clínicos se extraen textos médicos del dataset que posteriormente son procesados por los modelos de lenguaje natural. El objetivo es obtener el informe clínico final generado a partir de un informe clínico base, bien como resumen, diagnóstico o redacción asistida. Un resumen visual de estos pasos se muestra en [Figura 5.1](#). A continuación, se describen las etapas que se realizan para llevar a cabo dicha generación:

1. **Selección y preparación de los datos:** A partir del preprocesamiento y unificación realizados en la primera parte del proyecto, se han extraído distintos informes clínicos. Dependiendo del experimento, se han filtrado un pequeño subconjunto dependiendo de la categoría o han sido elegidos aleatoriamente. Por ejemplo, para la realización del wordcloud se han extraído textos en función del sistema afectado (circulatorio o respiratorio).
2. **Entrada al modelo de lenguaje:** El texto preprocesado se proporciona como contexto en la entrada (prompt) al modelo de lenguaje, formulado de la siguiente manera:

Based in <context>. <Instrucción>

En esta estructura, context respresenta el texto clínico, mientras que la instrucción define la tarea específica a realizar. Un ejemplo de instrucción puede ser: basado en el contexto, dime cuál es el diagnóstico para el paciente.

3. **Generación del informe:** El modelo produce una respuesta en función del prompt recibido, generalmente un texto nuevo que actúa como informe clínico generado. Este resultado puede ser un resumen, un informe completo o un texto de apoyo para el médico, dependiendo del prompt que le indiquemos.
4. **Evaluación del resultado mediante métricas:** Una vez generadas las respuestas, se aplican métricas para evaluar la calidad de generación del texto, entre otras, que son definidas en la siguiente sección.

5.2. Métricas de evaluación

Para evaluar la calidad de la generación de los informes por los modelos de lenguaje, se han utilizado métricas comunes en tareas de generación de texto. Estas métricas permiten

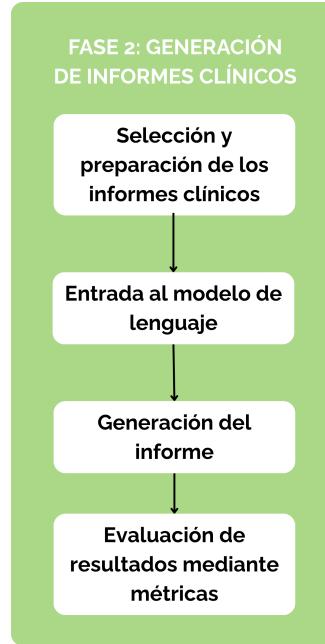


Figura 5.1.: Diagrama de flujo de trabajo de la generación automática de informes clínicos.

comparar los textos generados por cada uno de los modelos y proporcionan una estimación objetiva del rendimiento de la aplicación. Las métricas seleccionadas se describen en las siguientes subsecciones.

5.2.1. Métrica de calidad de generación del texto: BLEU

BLEU (Bilingual Evaluation Understudy) es una de las métricas automáticas más utilizadas para evaluar la calidad de generación de texto, especialmente en tareas de traducción automática. Introducido en [PRWZ02], BLEU compara el texto generado por el modelo con uno o varios textos de referencia humanos, calculando la coincidencia de n-gramas entre ambos.

Los n-gramas son secuencias consecutivas de n elementos, normalmente palabras o caracteres, extraídas de un texto. En el caso de palabras, un unígrafo es una sola palabra, un bigrama es una secuencia de dos palabras, y así sucesivamente. Por ejemplo, en la imagen Figura 5.2 vienen recogidos los diferentes n-gramas de la frase. Estos se utilizan comúnmente en procesamiento de lenguaje natural para modelar el contexto y la probabilidad de aparición de ciertas combinaciones de palabras. Cuanto mayor sea la superposición de n-gramas, mayor será la puntuación BLEU, que varía entre 0 y 1.

Otras métricas calculadas por BLEU son:

- **precision:** lista que contiene la precisión de n-gramas para distintos valores de n
- **brevity penalty:** penaliza hipótesis más cortas que la referencia
- **length ratio:** relación entre la longitud de la hipótesis y la de la referencia
- **translation length y reference length:** número total de palabras en la hipótesis y en la referencia, respectivamente

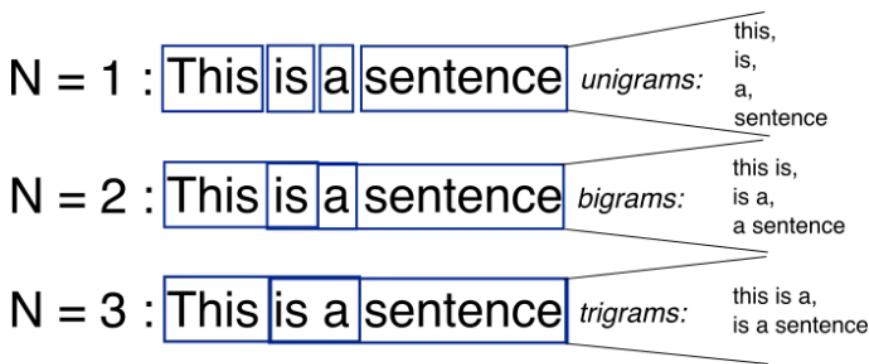


Figura 5.2.: Ejemplo de n-gramas. Fuente: <https://medium.com/@abhishhekjainindore24/n-grams-in-nlp-a7c05c1aff12>.

La métrica BLEU se calcula como

$$\text{BLEU} = \text{BP} \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right),$$

donde, BP es la brevity penalty, p_n es la precisión de los n-gramas, y w_n es el peso (normalmente uniforme).

5.2.2. Métrica de comparación de similitud de las respuestas: Bert score

BERTScore [ZKW⁺²⁰] es una métrica para evaluar la calidad de textos generados, basada en representaciones semánticas profundas extraídas mediante modelos de lenguaje preentrenados como BERT. A diferencia de métricas tradicionales como BLEU, que comparan la coincidencia de n-gramas, BERTScore calcula la similitud contextual entre cada token del texto generado y el de referencia, usando vectores embebidos. Para cada token, se busca su correspondencia más similar en el otro texto utilizando la similitud del coseno entre embeddings. Luego, se calcula una puntuación F1 a partir de las precisiones y recalls semánticas. La precisión mide cuántas de las predicciones positivas realizadas por el modelo son realmente correctas y el recall indica cuántos de los elementos positivos reales fueron correctamente identificados por el modelo. El score F1 es la media armónica entre precisión y recall, que proporciona una medida equilibrada cuando se desea un compromiso entre ambas.

Este enfoque permite capturar mejor la paráfrasis, sinonimia y reordenamientos gramaticales, ofreciendo una evaluación más cercana al juicio humano. Además, puede ajustarse mediante distintos modelos, como por ejemplo, uno entrenado para datos médicos.

5.2.3. Medidas de readabilidad y comprensión del texto

Evaluar la calidad de un texto no solo implica analizar su gramática o coherencia, sino también su facilidad de lectura, carga emocional y riqueza léxica. A continuación se describen algunas métricas comunes:

5. Generación automática de informes clínicos

Flesh Reading Ease Score Mide la facilidad de lectura del texto en inglés. Su fórmula se basa en la longitud de las oraciones y la cantidad de sílabas por palabra. Su rango se sitúa entre 0 y 100, cuanto más alto, más fácil de leer. Por ejemplo, 60 suele considerarse accesible para público general.

SMOG Index Estima el nivel educativo (en años) necesario para entender un texto, basado en la cantidad de palabras con tres o más sílabas. Cuanto más alto, más complejo.

Dale-Chall Readability Score Mide la legibilidad considerando solo palabras que aparecen en una lista de 3,000 palabras comunes del inglés. De esta manera, se penalizan los términos más difíciles. Su resultado se describe en puntos, los valores más bajos indican mayor facilidad. Por ejemplo, 8.0 implica un texto complejo.

Puntuación de polaridad de sentimientos Indica si el texto transmite un sentimiento positivo, negativo o neutro. El rango va desde -1 (muy negativo) a +1 (muy positivo).

Puntuación de subjetividad de sentimientos Mide cómo de subjetivo u objetivo es un texto. Su rango va desde 0 (objetivo) a 1 (completamente subjetivo). Por ejemplo, textos informativos como los informes clínicos deben tender a la objetividad.

Puntuación de diversidad léxica Calcula la variedad de vocabulario en un texto. Su rango también varía entre 0 y 1, siendo 1 una puntuación de diversidad léxica total.

Puntuación de riqueza léxica Evalúa la sofisticación y profundidad del vocabulario, considerando la frecuencia de palabras poco comunes o técnicas.

6. Experimentos y resultados

En este capítulo se explica el diseño de la experimentación, así como ejemplos reales de los experimentos y justificación de los modelos y datos utilizados. A continuación, se razonan las tecnologías utilizadas, aclarando el lenguaje de programación y librerías utilizadas. Finalmente, se analizan las métricas obtenidas al aplicar los modelos a los textos extraídos.

6.1. Diseño de la experimentación

Para realizar los distintos experimentos, primero se han realizado pruebas extrayendo una pequeña cantidad de textos, utilizándolos como entradas a los modelos. Más tarde se realizaron las pruebas definitivas, ejecutándolas en el servidor diseñado para ello. Por último, se han analizado las métricas obtenidas al realizar los experimentos. Las distintas pruebas realizadas se explican en las siguientes subsecciones.

6.1.1. Resumen de consulta clínica

El primer experimento tiene como objetivo verificar la capacidad de los modelos para resumir de forma clara y concisa. Para ellos se han realizado dos pruebas:

- **Prueba A (sin rol de especialista):** Se le proporciona al modelo un texto de entrada, y junto a él se le pide que devuelva un resumen del diagnóstico del paciente. De esta manera trabaja de forma generalista en la generación del resumen. El prompt empleado es el siguiente:

"role": "user", "content": "Based on context. Give me a summary of the diagnostic for the patient"

Siendo *context* el texto de entrada.

- **Prueba B (con rol de especialista):** Para esta prueba, además de pedirle al modelo que resuma el diagnóstico del paciente a partir del texto, también se le proporciona una instrucción de rol, en la que se especifica que actúe como un doctor profesional, proporcionando resúmenes claros y concisos. Esto permitirá estudiar si los modelos ponen más énfasis en la información relevante para un experto médico cuando redacta el resumen. El prompt utilizado es:

"role": "user", "content": "Based on context. Give me a summary of the diagnostic for the patient" , "role": "expert", "content": "'You are a professional medical doctor. You provide clear and concise summaries based on patient's clinical information'"

Siendo *context* el texto de entrada.

Estas pruebas han sido realizadas extrayendo los textos del servicio CARDIOTHORACIC. Para ello se han filtrado los textos y se han reducido a un tamaño de 150 textos de forma aleatoria. Para ambas pruebas se llevarán a cabo distintas métricas para evaluar el desempeño de estos modelos y comparar si determinar el rol puede influir en la generación de estos informes. Un ejemplo de input y output para esta prueba se encuentra en el Apéndice A.2.

6. Experimentos y resultados

6.1.2. Prueba de ventana de contexto

Para esta prueba también se le proporciona al modelo como contexto el historial médico de un paciente. A partir de él, se le pide que genere como salida un texto en formato JSON, en el que, en la primera línea escriba el diagnóstico del paciente, es decir, la enfermedad o problema que se describe que puede tener el paciente en dicha nota clínica. Para la siguiente línea se pide que describa la derivación del paciente, de manera que devuelva si el paciente fue enviado a hosp (hospital), ICU (Unidad de Cuidados Intensivos) o home (casa).

"role": "user", "content": "Based on context. In json format, write in a line: diagnostic; and what is the diagnostic for the patient. In other line write solution:, and if was derived to the hosp, to ICU or none "

Un ejemplo de input y output devuelto por el modelo se encuentra en [Sección A.3](#). Este procedimiento permite verificar la capacidad de comprensión de textos largos o densos para dar una salida más estructurada, de modo que extrae las características más importantes del resumen.

6.1.3. Justificación de la elección de los modelos

Se han utilizado principalmente tres modelos para la generación de estos informes clínicos. A continuación, se explica detalladamente cómo funciona cada uno de ellos y sus características principales.

6.1.3.1. Modelo gemma-2-2b-it

Gemma 2 2B es un modelo de lenguaje de código abierto desarrollado por Google DeepMind como parte de la familia Gemma 2 [[TRP⁺24](#)]. Con 2 mil millones de parámetros, Gemma 2 2B está diseñado para ofrecer un equilibrio entre rendimiento y eficiencia computacional, siendo especialmente adecuado para su ejecución en entornos con recursos limitados como CPUs modernas o GPUs de gama media.

Este modelo está basado en la arquitectura transformer, un modelo de aprendizaje profundo introducido en [[VSP⁺23](#)]. Su principal innovación es el mecanismo de autoatención (self-attention), que permite al modelo ponderar dinámicamente la importancia de cada palabra en una secuencia en relación con las demás, sin necesidad de procesar los datos en orden secuencial. Esto lo hace altamente paralelo y eficiente. Un Transformer típico se compone de capas repetidas de bloques de atención y capas feed-forward, acompañadas de mecanismos de normalización y conexiones residuales. Esta arquitectura ha demostrado una capacidad sobresaliente, y ha sido la base de casi todos los modelos de lenguaje avanzados actuales, como BERT, GPT, Llama o Gemma.

El entrenamiento de Gemma 2 se ha llevado a cabo sobre grandes volúmenes de datos en múltiples idiomas y dominios, garantizando una buena capacidad de generalización en tareas como generación de texto, comprensión de lenguaje natural y respuesta a preguntas.

Gemma 2 2B destaca por su licencia permisiva, lo que lo convierte en una alternativa viable para aplicaciones comerciales, académicas y personales.

En resumen, Gemma 2 2B representa una opción ligera pero potente dentro del ecosistema de modelos de lenguaje, ideal para desarrolladores y equipos que buscan integrar inteligencia artificial sin requerir infraestructuras de alto coste.

6.1.3.2. Modelo Llama 3.2-1B-Instruct

Llama 3.2-1B-Instruct es un modelo de lenguaje de la familia Llama 3 (Large Language Model Meta AI), desarrollado por Meta AI [GDJ⁺²⁴], y representa una de las variantes más compactas de esta serie, con aproximadamente 1.1 mil millones de parámetros. Esta versión Instruct está afinada específicamente para seguir instrucciones y mantener diálogos útiles, seguros y alineados con las intenciones del usuario, de manera que la hemos podido utilizar para generar informes de manera clara y concisa.

Gracias a su reducido tamaño, Llama 3.2-1B-Instruct es especialmente adecuado para su uso en dispositivos con recursos limitados, aplicaciones móviles o entornos donde la latencia y la eficiencia energética son críticas. Aunque más pequeño que sus hermanos mayores (como Llama 3 de 8B o 70B), conserva muchas de las mejoras introducidas en esta generación. Mantiene una arquitectura eficiente con atención causal unidireccional, capas normalizadas y técnicas modernas de entrenamiento que maximizan el rendimiento en tareas instruccionales, siguiendo el paradigma chat-instruct. Está disponible bajo una licencia abierta para investigación y desarrollo, lo que lo convierte en una opción viable para proyectos académicos y pruebas de concepto en producción. Este modelo demuestra que incluso con una escala reducida, es posible obtener un rendimiento competitivo en tareas conversacionales e instruccionales.

6.1.3.3. Modelo Qwen3-4B

Qwen3-4B es un modelo de lenguaje de propósito general desarrollado por Alibaba Group como parte de la tercera generación de su serie Qwen [YLY⁺²⁵], publicado en mayo de 2025. Con aproximadamente 4 mil millones de parámetros, Qwen3-4B se sitúa en la categoría de modelos de tamaño medio, ofreciendo un equilibrio entre capacidad de comprensión y generación de lenguaje, y eficiencia computacional. Está basado en una arquitectura Transformer decoder-only, optimizada con técnicas modernas como Grouped Query Attention (GQA), RMSNorm, y tokenización BPE multilingüe, lo que mejora su rendimiento en múltiples idiomas, incluidos el chino y el inglés. Esta versión fue preentrenada sobre un corpus mixto de alta calidad y luego afinada para mejorar su alineación con instrucciones humanas (Instruct tuning), lo que la hace especialmente útil en contextos conversacionales, generación de texto y asistencia interactiva. Además, se publica bajo una licencia abierta, que permite su uso en investigación y, bajo ciertas condiciones, en entornos comerciales. Su tamaño y diseño lo hacen especialmente adecuado para entornos de despliegue equilibrados entre rendimiento y coste.

6.1.4. Justificación de la elección de los conjuntos de datos

El dataset de este proyecto, MIMIC-IV, consta de millones de registros clínicos de pacientes. Al ser tan extenso y complejo, es difícil gestionarlo en toda su amplitud. Por ese motivo, hemos decidido limitarnos a un par de casos de uso completos, en lugar de intentar procesarlos en masa.

Principalmente, el procesamiento de toda la base de datos habría requerido una cantidad de recursos computacionales muy elevada, así como más tiempo destinado a la preparación de los datos como al análisis de los resultados de este proyecto.

Además, al centrarnos en unos pocos casos específicos, podemos estudiar en más profundidad cómo responde el modelo en contextos clínicos determinados. Así también es más

6. Experimentos y resultados

fácil detectar errores o incoherencias en los informes generados. Aún así, los casos elegidos están concebidos para ser representativos de diferentes especialidades. Por ejemplo, para los experimentos de resumen de consulta se han elegido los textos del servicio *cardiothoracic*. Esto proporciona una muestra suficiente para demostrar tanto las fortalezas como las debilidades de la aplicación en contextos más específicos.

Por tanto, el procedimiento de selección de casos específicos proporciona un equilibrio entre la profundidad en el análisis y una ejecución viable.

6.2. Tecnologías utilizadas

Para la realización de todos los pasos de programación de este proyecto se ha desarrollado código implementado en el lenguaje de programación Python. Este es un lenguaje de alto nivel de programación que destaca por la legibilidad de su código. Además, es el más utilizado en tareas de preprocesamiento, análisis de datos e inteligencia artificial, lo que proporciona una gran cantidad de recursos y librerías especializadas. Por ese motivo resulta la herramienta más adecuada para llevar a cabo la implementación del código. Toda la implementación queda recogida en el siguiente repositorio de GitHub: <https://github.com/ugritai/medicalReports>.

Dentro de Python, se han utilizado varias librerías de amplio uso en el ámbito. A continuación se describen las más relevantes para el proyecto:

- **Pandas:** Librería para manipular, analizar y limpiar datos en formas de tablas (DataFrames). Se ha utilizado para todo el preprocesamiento, así como para manejar y manipular los informes clínicos y los resultados de los modelos. La estructura de datos de Pandas proporciona una forma eficaz e intuitiva de llevar a cabo operaciones de procesamiento de datos, lo cual era necesario para gestionar el dataset.
- **Dask:** Framework de datos en paralelo o distribuido, ampliando pandas o numpy a más memoria o más núcleos. En algunos casos se ha manejado para leer y procesar algunos archivos del dataset de manera más eficiente que pandas.
- **Bokeh:** Herramienta de visualización interactiva para crear gráficas en la web directamente desde Python. Todas las gráficas del proyecto han sido generadas con esta librería.
- **Transformers:** Librería de Hugging Face para implementar modelos de procesamiento de lenguaje natural. Se ha utilizado para cargar los modelos de procesamiento de texto.
- **Torch:** PyTorch es la base de muchos modelos de deep learning, usada para procesamiento de texto. Se ha utilizado como framework de bajo nivel en el que están implementados los modelos de transformers.

El código ha sido desarrollado a través de un ordenador personal, pero debido a la complejidad de algunos procesos, se han utilizado dos servidores para realizar el proyecto. Estos dos servidores se han utilizado para diferentes tareas:

- **Servidor Asus RS720 - 2x Xeon 4210R - 128GB DDR4 2666 - 2xRTX A5000 - 2x1TB - 10TB SATA3:** Este servidor de alto rendimiento está equipado con 2 GPUs NVIDIA RTX A5000, 2x CPUs Intel Xeon Silver 4210R, 128 GB de RAM y con amplio almacenamiento. Proporciona la potencia de cálculo necesario para entrenar y ejecutar modelos

de generación de texto. Este equipo es indispensable para aplicar inteligencia artificial para tareas de procesamiento de lenguaje natural, como la generación, traducción o resumen de textos clínicos. Por tanto, ha sido utilizado para la ejecución de pruebas en los distintos modelos mencionados anteriormente.

- **Servidor Asus RS700-E9-RS4 / 2 x Xeon Silver 4114 10x2,2 Ghz / 128 GB DDR4 2666 / 2x4TB / 10 GBe:** Este segundo servidor está destinado al procesamiento de datos en paralelo o distribuido. Está equipado con 2x CPUs Intel Xeon Silver 4114, 128 GB de RAM, 10GBE de red y tiene una gran capacidad de almacenamiento en disco. Esto proporciona una plataforma estable y capaz de manipular grandes cantidades de información. Por este motivo resulta idóneo para preprocesar, unificar y analizar del dataset del proyecto. También se ha utilizado para preparar los textos clínicos para la entrada a los modelos de IA y llevar a cabo otras tareas dentro de procesamiento.

Por último, cabe destacar que los textos clínicos extraídos del dataset están redactados en el idioma inglés. Por esta razón se han utilizado tecnologías en este idioma, principalmente modelos de Hugging Face que están optimizados para procesar textos en dicho idioma, ya que han sido entrenados en este lenguaje.

6.3. Análisis de métricas

6.3.1. Longitud media de los textos

Una de las métricas más importantes de análisis de textos que no ha sido mencionada hasta ahora es la longitud. Es importante saber cuán largo es un texto, sobre todo a la hora de extraer resúmenes.

En la [Tabla 6.1](#) se recogen las medias de la longitud de palabras de todos los textos, tanto de entrada como de salida. En la primera línea observamos que la longitud media de los textos de entrada es 1440.46 palabras, lo cual demuestra que los informes clínicos extraídos están cargados de información.

Sin embargo, cuando se generan los resúmenes más concisos con los modelos, la longitud de las salidas se reduce drásticamente. Esto tiene sentido, ya que la tarea planteada es resumir la información para obtener textos más ágiles y fáciles de leer, sin perder sus características más relevantes.

A partir de los resultados podemos destacar que Gemma produce textos muy concisos con alrededor de 95 palabras en la mayoría de casos, sin una variabilidad importante en la longitud cuando varía el rol o la temperatura. El modelo Llama muestra más variabilidad en la longitud que Gemma, descendiendo en los modelos con rol respecto a los sin rol. Esto indica que dar al modelo un rol más específico puede llevar a una mayor concisión, probablemente porque así logra centrarse en los hechos más importantes. Qwen supera la longitud con respecto a los demás modelos, con alrededor de 570 palabras. Además, el rol tiene poca influencia en este modelo, lo que podría interpretarse como que el modelo tiene una mejor capacidad de resumen y no necesita tanto contexto de rol.

Por último, en todos los casos podemos notar que la reducción de longitud es muy alta. Esto evidencia que la tarea de resumen clínico logra eliminar información secundaria o redundante, conservando así lo más importante.

6. Experimentos y resultados

Modelo	Temperatura	Rol	Longitud (palabras)
Input	.	.	1440.46
	1.0	Sin rol	95.02
		Con rol	94.97
	Gemma	0.5	95.31
		Con rol	94.72
		Sin rol	95.19
		0.1	94.6
		Con rol	95.52
		1.0	80.43
		Sin rol	85.68
Llama	0.5	Con rol	79.17
	0.1	Sin rol	87.4
		Con rol	81.4
	Qwen	1.0	576.41
		Con rol	566.7
		Sin rol	581.32
Qwen	0.5	Con rol	579.56
	0.1	Sin rol	578.39
		Con rol	586.69

Tabla 6.1.: Comparación de la longitud del texto de entrada y las salidas de los modelos de generación de texto según su temperatura y rol.

6.3.2. Métrica BLEU

Para analizar esta métrica se necesitan clasificar los textos en predicciones y referencias. Se ha utilizado para cada texto:

- **Predicciones:** texto de salida generado por el modelo, es decir, el resumen clínico generado.
- **Referencias:** su texto correspondiente extraído del dataset. Como no tenemos referencias de expertos que realicen el resumen, se usa el texto original del dataset para ver cómo se comportan los distintos modelos entre sí, en lugar de ver si alcanza el valor del resumen experto.

En las [Figura 6.1](#), [Figura 6.2](#) y [Figura 6.3](#) podemos observar que los valores BLEU son extremadamente bajos. Esto es debido a que se compara como hipótesis el texto resumido contra el texto de referencia completo. De este modo, la disparidad en longitud y contenido resulta en estas puntuaciones tan bajas.

A continuación se comparan los valores BLEU de cada modelo, pero si nos damos cuenta, al ser tan bajos, la diferencia es muy sutil, ya que la puntuación es prácticamente 0. Para Gemma, los valores sin rol son consistentemente más altos que con rol. La temperatura 0.5 muestra el mejor rendimiento. Llama realiza un comportamiento opuesto a Gemma, superando significativamente la prueba con rol a sin rol. Existe una mayor variabilidad, pero la temperatura 0.5 es nuevamente la óptima.

En el caso de Qwen, los valores BLEU son sustancialmente más altos (0.04-0.05), sin ser tan cercanos a 0. Las diferencias entre rol y sin rol son mínimas, y el rendimiento es ligeramente

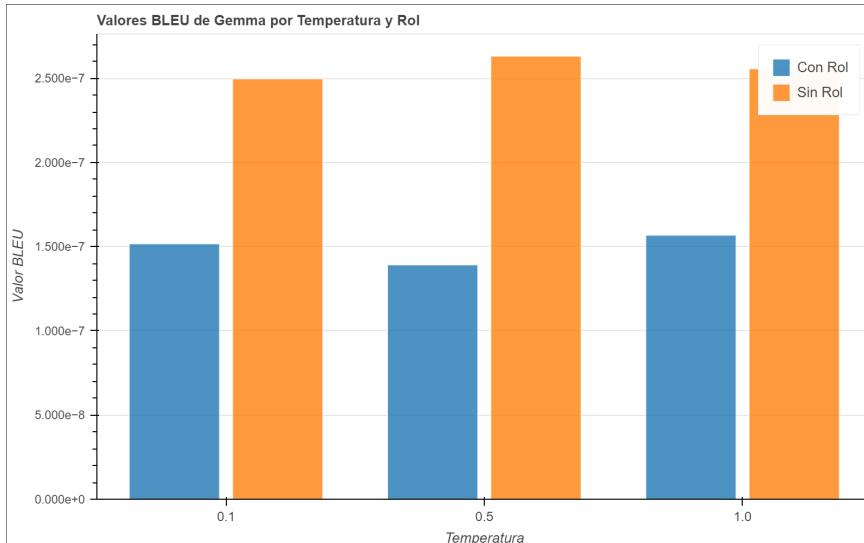


Figura 6.1.: Valores BLEU de Gemma por temperatura y rol.

decreciente con el aumento de temperatura.

Como conclusión para esta métrica, podemos decir que Qwen muestra el mejor rendimiento, con valores de magnitud superiores. Esto podría ser debido a la longitud de los textos generados, comentada en el apartado anterior. La temperatura 0.5 parece ser la óptima para la mayoría de configuraciones.

6.3.3. Métrica de comparación de similitud de las respuestas: Bert score

En la Figura 6.4, Figura 6.5 y Figura 6.6 se muestra una comparación entre dos variantes de BERTScore: el modelo BERT estándar ¹ y BioClinical_BERT ², especializado en el dominio médico. En ellas se compara el rendimiento de los modelos bajo distintas temperaturas, a través de la precisión, recall y F1. La hipótesis se construye a partir de la respuesta del modelo sin rol, mientras que la referencia es la respuesta del mismo modelo con la misma temperatura con rol específico.

Todas las métricas muestran un rendimiento muy alto y una notable estabilidad a través de las diferentes temperaturas, lo que contrasta el comportamiento de la métrica BLEU. Esto sugiere que BERTScore es más robusta ante variaciones en el texto generado. Generalmente, en cuanto al rendimiento, Qwen es el que mejor resultados obtiene, entre 0.87 y 0.89. Le sigue Gemma con 0.85-0.89 y, por último Llama con 0.72-0.8.

Analizando por métrica podemos decir que:

- **Precision:** Es la métrica que más variabilidad presenta. Por ejemplo, BioClinical_BERT obtiene resultados muy parecidos para Gemma y Qwen, mientras que Llama es la que peor puntuación obtiene. En cambio, para BERT por defecto, los valores de Gemma y Qwen varían más, y Llama no queda tan alejada de los otros dos. Qwen muestra la precisión más alta, sugeriendo que podría ser menos propenso a generar información irrelevante.

¹<https://huggingface.co/distilbert/distilbert-base-uncased>

²https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

6. Experimentos y resultados

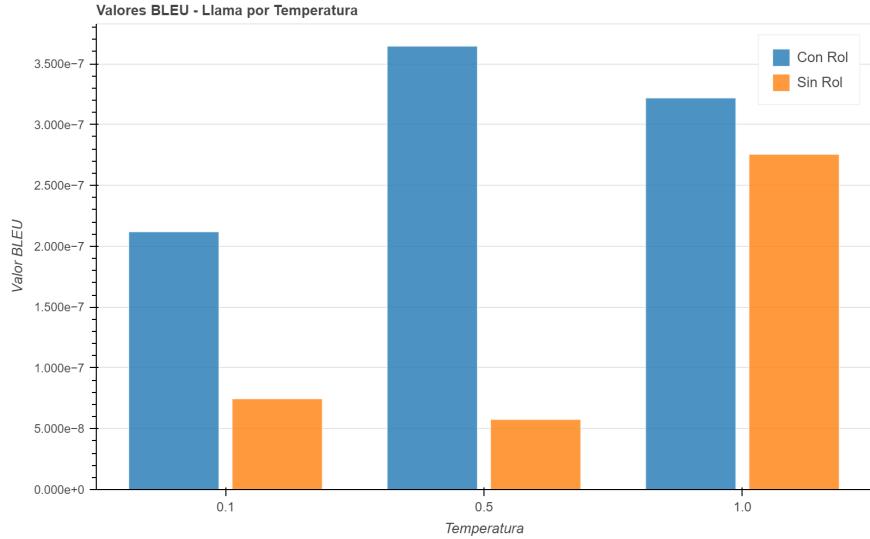


Figura 6.2.: Valores BLEU de Llama por temperatura y rol.

- **Recall:** En esta gráfica, los valores de Gemma y Qwen mejoran ligeramente para el modelo BioClinical_BERT, aunque para Llama no lo hacen. Para Bert por defecto, se obtienen resultados parecidos, pero para Llama mejoran en 0.3 puntos. De nuevo Qwen logra el recall más alto y Llama el más bajo, lo que significa que podría omitir más detalles relevantes.
- **F1:** La media entre las dos métricas anteriores sigue mostrando resultados muy parecidos aunque se modifique la temperatura, en la que destaca el modelo Gemma con BERT por defecto, y Qwen con BioClinical_BERT. La puntuación F1 devuelve valores generalmente altos, lo que indica un buen rendimiento general de los modelos.

Comparando los dos modelos de BERT, podemos observar que en BERT por defecto se obtiene una puntuación más alta que el otro modelo en prácticamente todos los casos. Aunque parezca que el modelo especializado debería dar mejores resultados para textos clínicos, esto no debería ser así. El modelo clínico, al estar especializado, tiene un vocabulario y estilo muy específicos. De esta manera, cuando comparamos nuestros textos con el modelo por defecto, obtiene mejor puntuación que el especializado, porque este último penaliza un vocabulario más general, ya que entiende más sobre ese tema en cuestión.

En cuanto a la temperatura, no influye en estos resultados, ya que no varían al modificarla. El rendimiento en todas las métricas se mantiene estable cuando la temperatura aumenta de 0.1 a 1.0. Esto indica que se pueden fomentar salidas más diversas de los modelos sin sacrificar la calidad o la consistencia de los textos. Incluso para temperaturas con una mayor "creatividad", siguen generando resúmenes relativamente precisos.

En conclusión, los modelos que obtienen mejor rendimiento son Gemma y Qwen. En las pruebas realizadas con BERT por defecto, Gemma obtiene mejor puntuación, mientras que para BioClinical_BERT, destaca Qwen. No obstante, la diferencia entre Gemma y Qwen es prácticamente nula, pero Llama obtiene peores resultados. Sin embargo, si tuviéramos que elegir qué modelo tiene el mejor rendimiento sería Qwen, ya que es el que mejor resultados obtiene en la mayoría de métricas. Queda demostrado que la metodología de comparar las

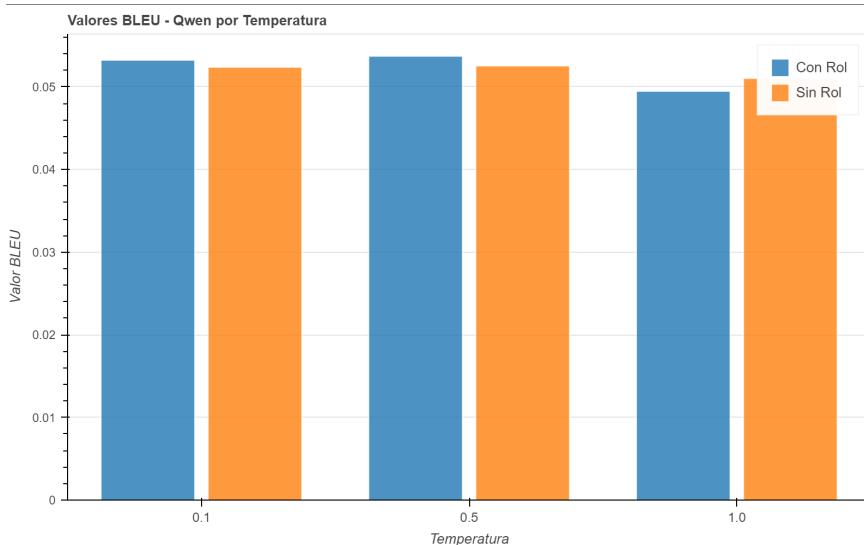


Figura 6.3.: Valores BLEU de Qwen por temperatura y rol.

respuestas de los modelos con especificación de rol mide qué tan bien retiene la información cuando es guiado por un rol, ofreciendo puntuaciones altas.

6.3.4. Medidas dereadabilidad y comprensión del texto

En esta subsección, se evalúa la calidad y la comprensión de los diferentes textos clínicos generados por los modelos, considerando su comportamiento con rol y sin rol, y bajo diferentes temperaturas.

Flesch Reading Ease En la [Figura 6.7](#) se evalúa el índice Flesch Reading Ease. Llama (con rol) es el modelo que produce los textos más fáciles de leer, alcanzando puntuaciones superiores a 50 y acercándose a 60 en algunos casos. Qwen también genera textos con buena legibilidad, especialmente sin rol, con puntuaciones por encima de 40. Gemma tiene la legibilidad más baja, especialmente con rol, con puntuaciones que a menudo están por debajo de 35, lo que sugiere que sus resúmenes son más difíciles de leer. El impacto del rol varía significativamente entre los modelos: para Gemma y Qwen, los textos sin rol son más fáciles de leer que los textos con rol, mientras que para Llama, los textos con rol son notablemente más fáciles de leer. La modificación de la temperatura generalmente no varía los resultados, pero por ejemplo, en el caso de Llama con rol sí que se ve una clara disminución cuando aumenta la temperatura.

En esta métrica, Llama destaca por generar los resúmenes más fáciles, lo cual es ventajoso para asegurar la comprensión por parte de los profesionales y los pacientes. Por otro lado, Gemma produce textos que son más difíciles de leer.

Smog Index Este índice estima el nivel educativo necesario para comprender un texto en la [Figura 6.8](#). Para Gemma y Qwen, los textos sin rol suelen tener un índice ligeramente más bajo. Para Llama, los textos con rol tienen un índice significativamente más bajo, de manera

6. Experimentos y resultados

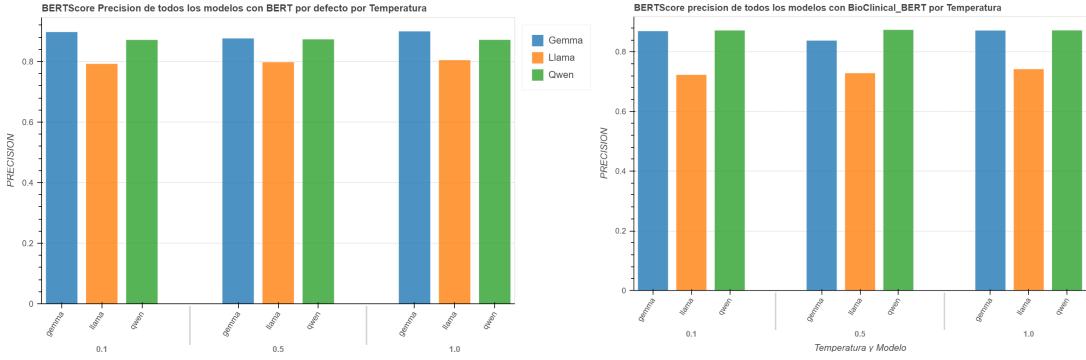


Figura 6.4.: Valores de precisión para BERTScore para todos los modelos por temperatura. A la izquierda se muestran los valores utilizando el modelo BERT por defecto para el BERTScore. A la derecha se muestran los valores utilizando el modelo BioClinical_BERT.

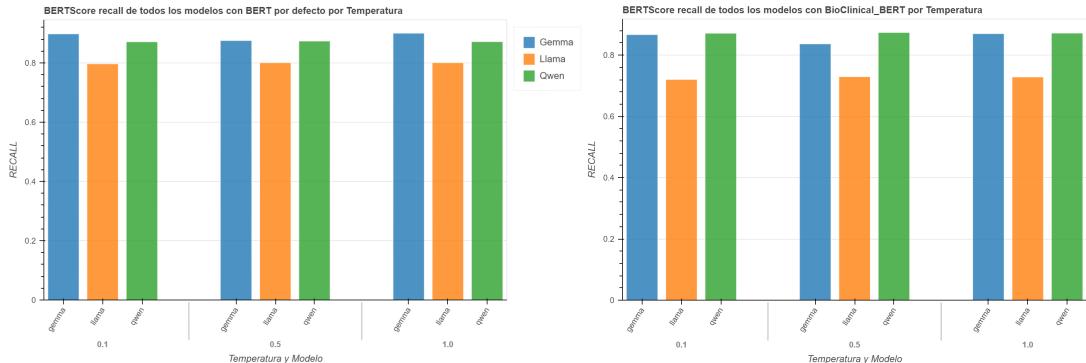


Figura 6.5.: Valores de recall para BERTScore para todos los modelos por temperatura. A la izquierda se muestran los valores utilizando el modelo BERT por defecto para el BERTScore. A la derecha se muestran los valores utilizando el modelo BioClinical_BERT.

que son más fáciles de entender. En cuanto a la temperatura, no se observa una tendencia clara de aumento o disminución con la temperatura.

Como conclusión, Llama, especialmente con rol, genera los textos más comprensibles, lo cual es consistente con su buena puntuación en Flesch Reading Ease. Gemma por el contrario, produce los textos más complejos de entender, necesitando un nivel educativo mayor para poder comprenderlos.

Dale Chall Index El índice Dale Chall es otra medida que enfoca en el porcentaje de palabras difíciles. En la Figura 6.9 el impacto del rol sigue siendo variable. Para Qwen y Gemma, este valor no varía mucho, pero para Llama, con rol se obtienen valores más altos para esta métrica, lo que sugiere que los textos son más sencillos de leer con esta configuración.. Las puntuaciones se mantienen relativamente estables con la temperatura, sin cambiar notablemente en las mismas configuraciones de los modelos.

Llama sigue siendo el mejor en legibilidad. La alta puntuación de Qwen en este índice

6.3. Análisis de métricas

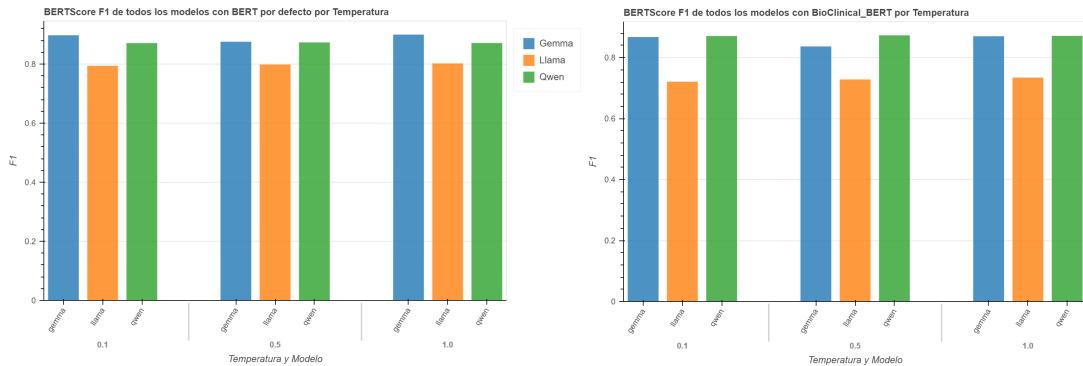


Figura 6.6.: Valores de F1 para BERTScore para todos los modelos por temperatura. A la izquierda se muestran los valores utilizando el modelo BERT por defecto para el BERTScore. A la derecha se muestran los valores utilizando el modelo BioClinical_BERT.

podría deberse a un uso más frecuente de palabras técnicas o menos comunes en los textos generados.

Puntuación de polaridad de sentimientos La polaridad de sentimientos mide si el sentimiento expresado en un texto es positivo, negativo o neutro. En la figura Figura 6.10 podemos observar que pueden parecer cambios más drásticos en esta puntuación, pero apenas varían 2 décimas. En este caso, el impacto del rol y la temperatura se ve mayor reflejado. Gemma con rol muestra una polaridad consistentemente positiva, que tiende a aumentar con la temperatura. Esto sugiere que al operar con rol, Gemma tiende a usar un lenguaje con una carga emocional más fuerte. Qwen con rol obtiene resultados parecidos a Gemma, pero con una polaridad ligeramente menor. Sin embargo, Llama con rol presenta una polaridad consistentemente negativa, que se vuelve más negativa a temperaturas más altas. Por último, las versiones sin rol de todos los modelos tienden a ser más neutrales.

Puntuación de subjetividad de sentimientos Para esta métrica, en la Figura 6.11, podemos analizar que para todos los modelos, la puntuación de subjetividad es generalmente más alta cuando el modelo opera con rol en comparación con sin rol. Esto es un hallazgo intrigante, que podría interpretarse como que al adoptar un rol específico, los modelos incorporan un lenguaje que, aunque intenta ser profesional, es percibido como más subjetivo por el algoritmo de análisis. Aún así, con Llama sin rol obtiene mayores valores que con rol.

El impacto de la temperatura se ve reflejado en esta métrica, aumentando los valores de subjetividad conforme aumenta la temperatura. Esto sugiere que a mayor creatividad, se vuelven más subjetivas las respuestas.

Puntuación de diversidad léxica En la Figura 6.12 se analiza la diversidad léxica, que mide la variedad de palabras distintas utilizadas en un texto. El impacto del rol apenas varía la puntuación en este caso, excepto para Llama, que con rol obtiene mayor diversidad léxica que sin rol. Respecto al impacto de la temperatura, esta métrica parece ser bastante robusta a estos cambios para todos los modelos.

6. Experimentos y resultados

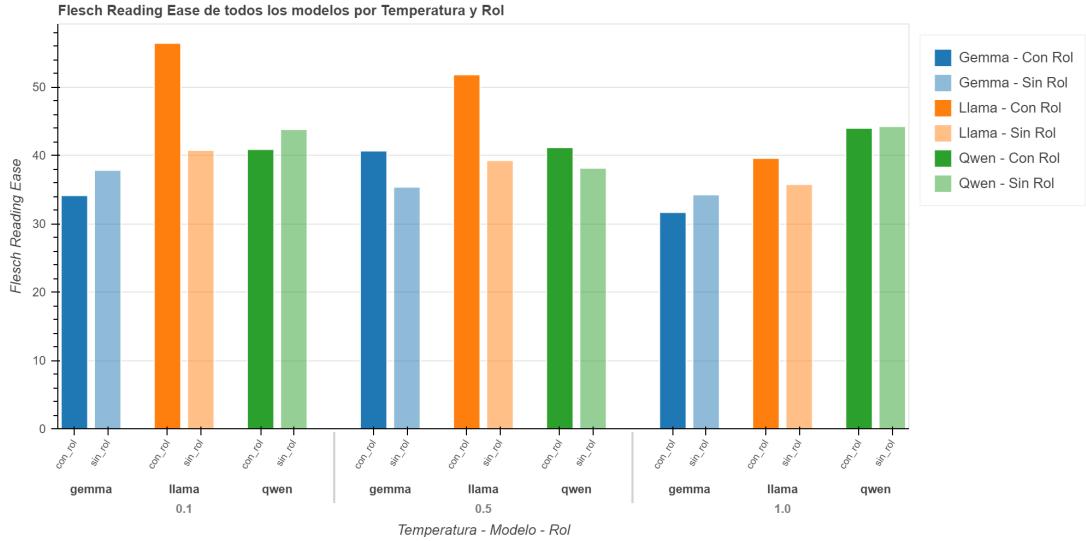


Figura 6.7.: Puntuación Flesch Reading Ease de todos los modelos por temperatura y rol.

Llama es el modelo que genera textos con mayor diversidad léxica, lo cual es positivo para la riqueza del lenguaje en los resúmenes.

Puntuación de riqueza léxica La riqueza léxica, en la Figura 6.13, es una medida más estricta de la variedad de palabras, considerando la proporción de palabras únicas respecto al total de palabras. En este caso, existe una clara división: Gemma y Llama muestran una riqueza léxica significativamente más alta en comparación con Qwen. Para los valores de todos los modelos, el impacto del rol apenas se ve reflejado, variando solamente unas centésimas. Además, el impacto de la temperatura tampoco se ve reflejado, manteniéndose estable a través de las diferentes temperaturas.

Para esta métrica, Gemma y Llama sobresalen en riqueza léxica, lo que implica que sus resúmenes son léxicamente más densos y variados. La baja riqueza léxica de Qwen podría resultar en textos más monótonos o con información redundante.

Como conclusión, podemos comentar varios aspectos sobre estas métricas. El impacto del uso de rol influye significativamente en la subjetividad, la diversidad léxica y la polaridad de sentimientos, de manera variada entre los modelos. A veces, mejora algunas métricas y otras las empeora, pero en general el uso de rol influencia las respuestas de los modelos. En cambio, la temperatura no tiene un impacto tan drástico como podría esperarse. Las tendencias son más sutiles y varían entre modelos y métricas.

6.3.5. Consideraciones finales

Tras analizar en profundidad las métricas de evaluación recogidas, podemos destacar varias consideraciones en relación con el comportamiento de los modelos, la influencia de parámetros de rol y la temperatura en la generación de textos.

En primer lugar, el uso de rol en el prompt tiene un impacto apreciable en varias métricas, en particular la subjetividad, la diversidad léxica y la polaridad de sentimientos. Esto pone de

6.3. Análisis de métricas

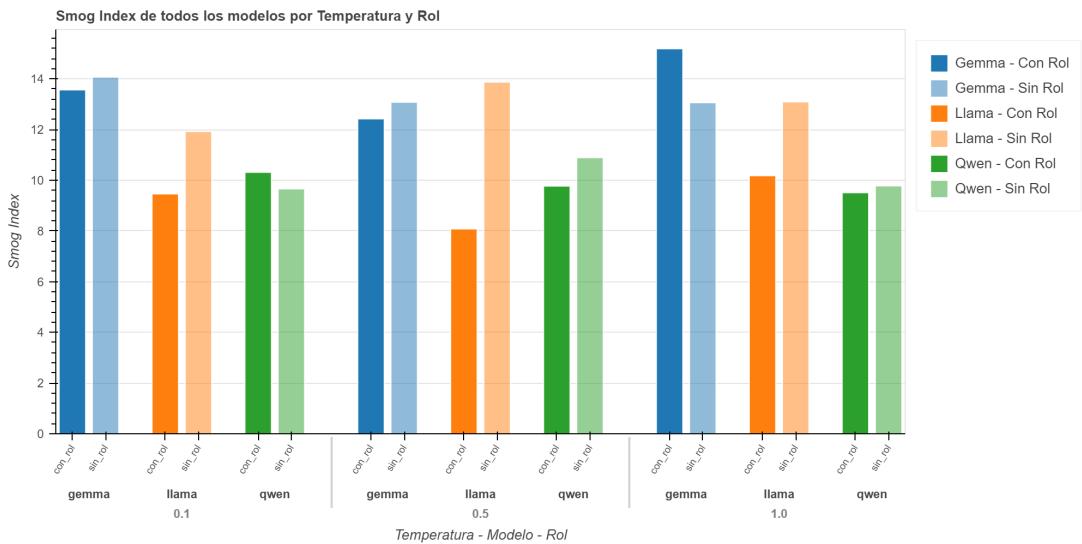


Figura 6.8.: Indice Smog de todos los modelos por temperatura y rol.

manifiesto que dar al modelo un rol más específico puede llevar tanto a textos más concisos como más subjetivos, pero también tiene consecuencias en la riqueza léxica.

En segundo lugar, la temperatura tiene en general un impacto más sutil en las métricas. Esto pone de manifiesto que el modelo es capaz de producir textos relativamente estables en distintas temperaturas, sin sacrificar, en general, la claridad o riqueza de los textos.

Como conclusiones finales de cada modelo, podemos destacar que:

- **Qwen:** Muestra la mayor precisión en BERTScore. Proporciona texto más específicos y más estables pero más monotonos, es decir, con menor diversidad y riqueza léxica. Además, obtiene la mayor longitud media de palabras en diferencia a los demás modelos.
- **Llama:** Produce textos más fáciles de leer, más variados, pero más emotivos cuando reciben rol. Sin embargo, es el modelo que peor puntuación recibe en la mayoría de métricas, sobre todo en BERTScore.
- **Gemma:** Logra textos más ricos en vocabulario pero más subjetivos o más difíciles de leer. Se sitúa en varias métricas muy cerca de Qwen, superándole en algunas.

Este análisis evidencia que cada modelo tiene fortalezas y debilidades, y que tanto el rol como la temperatura están relacionados con el estilo de generación. Si tuviéramos que elegir un modelo entre los tres, Qwen es el que destaca en la mayoría de métricas, obteniendo puntuaciones excelentes. En cambio, si buscamos resúmenes no tan extensos, podríamos elegir Gemma, que se sitúa también en puntuaciones altas. En cuanto a la especificación de rol, la alternativa con rol es la que obtiene mejores resultados.

6. Experimentos y resultados

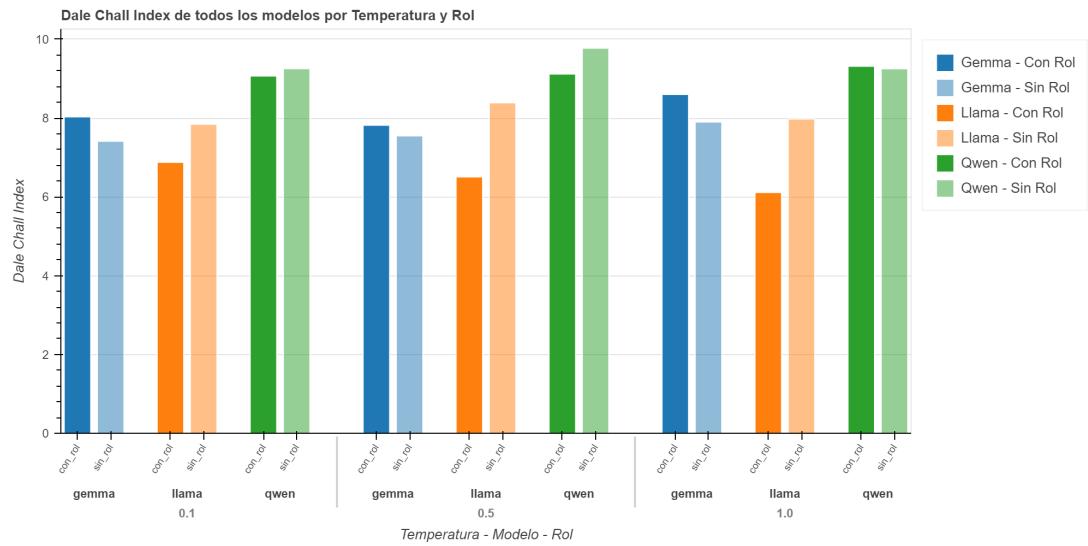


Figura 6.9.: Indice Dale Chall de todos los modelos por temperatura y rol.

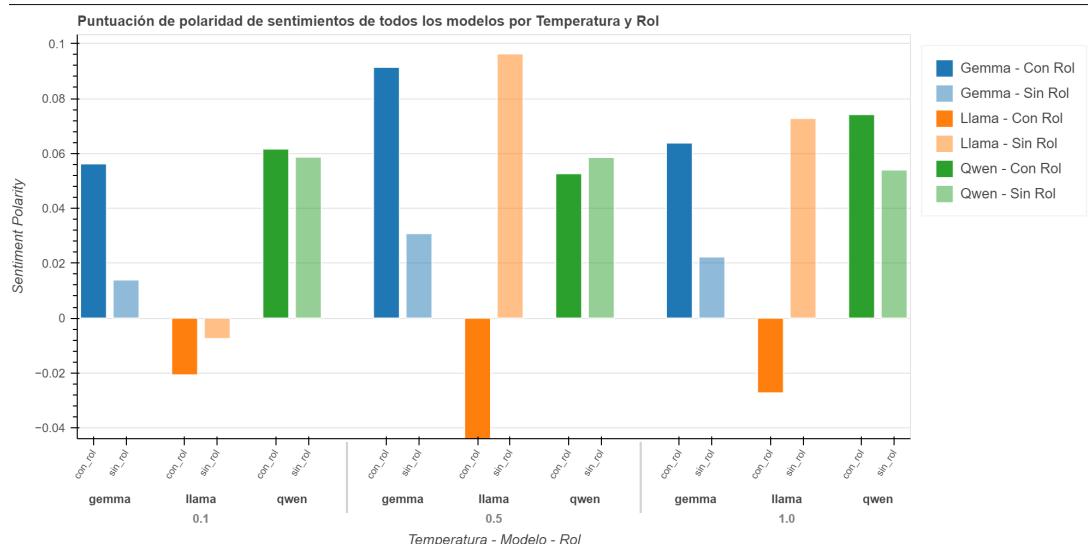


Figura 6.10.: Indice de polaridad de sentimientos de todos los modelos por temperatura y rol.

6.3. Análisis de métricas

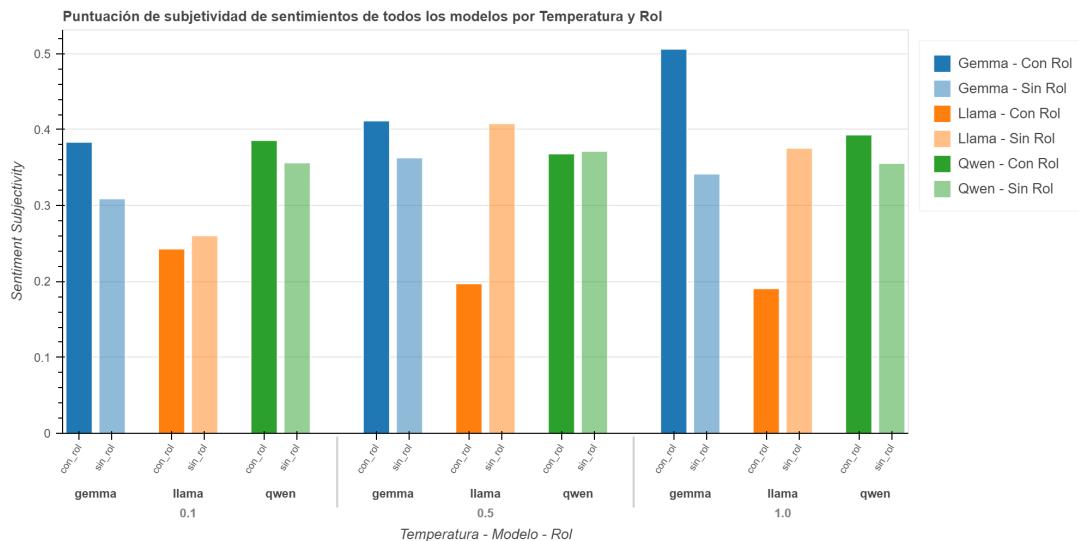


Figura 6.11.: Indice de subjetividad de sentimientos de todos los modelos por temperatura y rol.

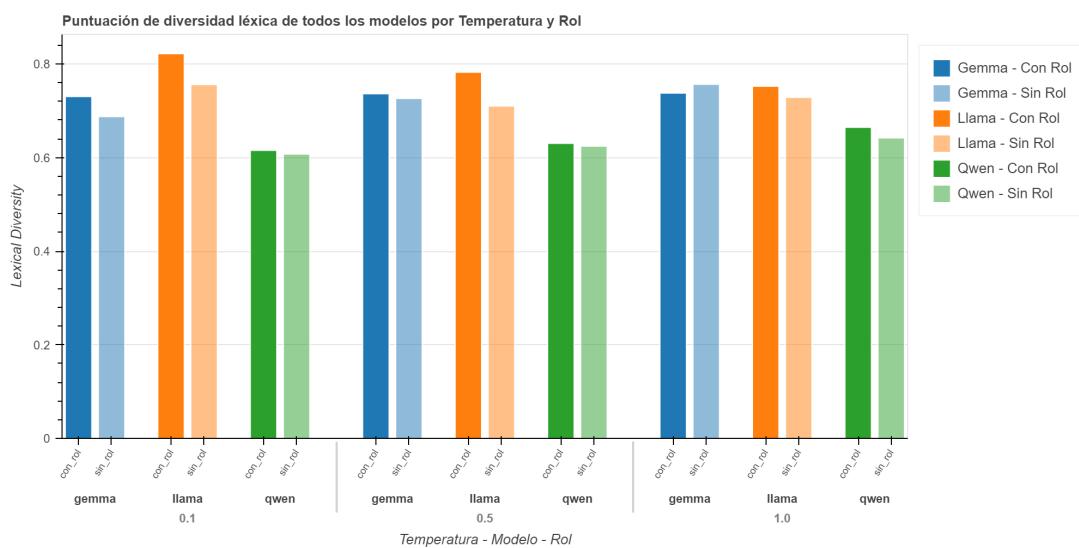


Figura 6.12.: Puntuación de diversidad léxica de todos los modelos por temperatura y rol.

6. Experimentos y resultados

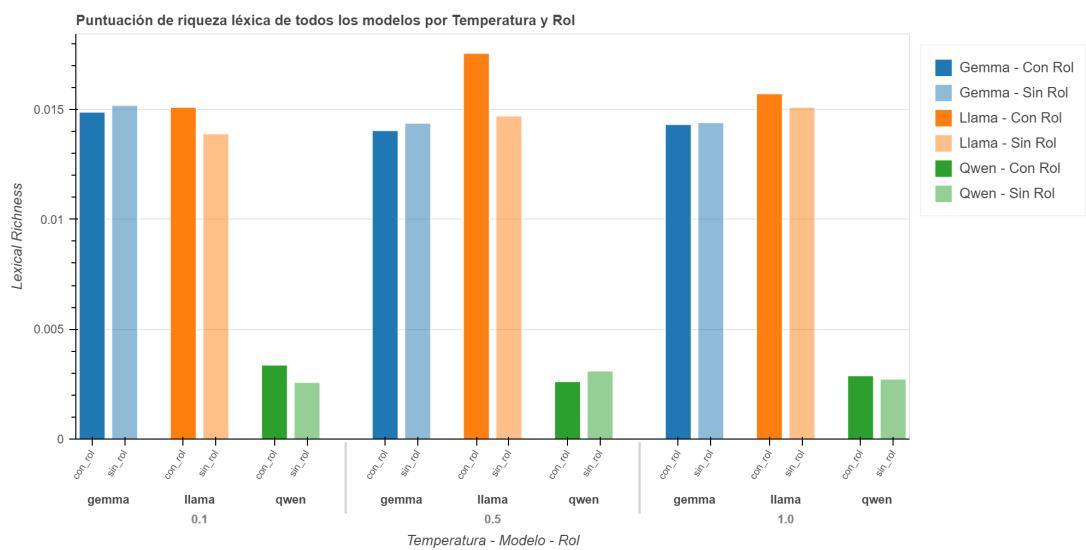


Figura 6.13.: Puntuación de riqueza léxica de todos los modelos por temperatura y rol.

7. Conclusiones y trabajo futuro

En este capítulo final, se razona si se han alcanzado los objetivos planteados en el proyecto, recogidos en las conclusiones, y, por último, se proponen trabajos futuros.

7.1. Conclusiones

Este Trabajo de Fin de Grado tiene como objetivo general desarrollar una aplicación capaz de generar automáticamente informes clínicos a partir de textos clínicos, ayudando así a agilizar y estandarizar el procedimiento de redacción de informes en entornos hospitalarios. El problema planteado era la sobrecarga de tareas administrativas, en particular en lo relacionado con la redacción manual de informes, que consume una parte importante del tiempo de los profesionales sanitarios y puede dar lugar tanto a inconsistencias como a errores en el manejo de la información clínica. Además, incluye que la cantidad de información clínica supera la capacidad humana, y los sistemas actuales ofrecen poca automatización de los historiales clínicos.

Podemos destacar que el proyecto ha alcanzado con éxito el objetivo planteado, demostrando así que es posible aplicar modelos de inteligencia artificial generativa en el entorno médico para resumir textos clínicos, destacar la información más relevante y producir informes más claros y concisos. Por tanto, podemos realizar una valoración positiva de los resultados, ya que el modelo proporciona una base para aliviar parte de la carga de los trabajadores del ámbito de la salud.

A lo largo de este procedimiento se han alcanzado los objetivos específicos planteados al inicio:

- **OE1.** Estudiar el estado del arte en generación de informes clínicos. Esto se ha realizado en el capítulo 2, donde se analiza en profundidad el uso de IA en el ámbito de la salud, los modelos de generación de texto más avanzados y la aplicación del dataset MIMIC-IV en la investigación.
- **OE2.** Analizar, limpiar y preparar el conjunto de datos MIMIC-IV. La ejecución de este objetivo puede apreciarse en el capítulo 4, tanto en el procedimiento de preparación de los datos, como en el análisis exploratorio de datos.
- **OE3.** Implementar un modelo de IA capaz de generar textos clínicos. Esto se ha llevado a la práctica en el capítulo 5, en el procedimiento de generación de informes clínicos.
- **OE4.** Evaluar los modelos realizando experimentos. Finalmente, en el capítulo 6, en el diseño de la experimentación y el análisis de los resultados, se muestran las métricas de evaluación y se comparan las pruebas realizadas.

A nivel personal, la realización de este proyecto me ha supuesto una experiencia muy enriquecedora tanto en el plano técnico como en el formativo. Durante el ejecución de este trabajo he tenido la oportunidad de familiarizarme con distintos aspectos específicos que

7. Conclusiones y trabajo futuro

no había profundizado en el desarrollo del grado, especialmente, la ciencia de datos y el uso de LLMs. Por ejemplo, he aprendido a manipular y gestionar grandes volúmenes de información con la librería Pandas, desde la selección de datos hasta la preparación de textos clínicos para dar como entrada a los modelos de generación de texto.

Además, he utilizado LLMs, una herramienta compleja pero muy relevante en el campo de la IA. He aprendido algunos aspectos éticos sobre el manejo de datos médicos de distintos pacientes, tomando conciencia de la importancia de respetar la privacidad de los pacientes. Por esto, el procedimiento de anonimización y manejo de datos ha ocupado una parte importante en el desarrollo.

En definitiva, me siento muy orgulloso con el resultado de este trabajo, ya que me ha servido tanto para ampliar mis habilidades en el manejo de datos, en el uso de modelos de IA, en el procesamiento de textos clínicos, así como en la redacción de documentos o informes dentro del campo de la investigación. Además, me ha dado la oportunidad de participar en un proyecto tutelado, permitiéndome trabajar en equipo y contrastar diferentes puntos de vista. Esto, sin duda, constituye una base muy valiosa para futuros trabajos y para mi desarrollo profesional.

Este proyecto proporciona así una base tanto teórica como práctica para futuros desarrollos en el ámbito de la generación de informes clínicos con IA, demostrando que es posible aplicar modelos de última generación en entornos hospitalarios, siempre que se tenga en cuenta la complejidad de los textos. Además, el conjunto de datos MIMIC-IV resulta clave en este proyecto, ya que proporciona una fuente de información realista y de calidad.

7.2. Trabajos futuros

A partir de los resultados alcanzados, se puede continuar esta línea de investigación realizando:

- **Ampliar el conjunto de datos:** Por limitaciones de capacidad de procesamiento, en el proyecto se trabajó con una pequeña muestra de MIMIC-IV. Un paso para futuros desarrollos es aplicar el procedimiento a todo el dataset, aumentando así tanto la robustez como la generalización de los modelos.
- **Aumentar la variedad de textos clínicos:** Actualmente el modelo proporciona resúmenes de los diagnósticos, pero se podría ir más lejos generando información como prescripciones, pruebas de laboratorio o seguimiento del paciente. Esto podría dar lugar a informes más completos.
- **Explorar otros modelos de IA:** Aumentar el rango de modelos usados, así como considerar nuevos métodos de generación específicos para textos clínicos, como por ejemplo, realizar un fine-tuning.
- **Aplicar otras métricas de generación de texto:** Además de las métricas utilizadas en el proyecto, podría resultar conveniente aplicar otras métricas más avanzadas, con el objetivo de realizar una evaluación más exhaustiva. En este caso, deberíamos haber analizado la métrica Self-BLEU, en lugar de BLEU, ya que en esta primera se compara la diversidad de un texto generado contra todas las frases generadas en un conjunto.
- **Desarrollar una interfaz intuitiva:** Finalmente, como paso a la aplicación práctica de esta herramienta, sería conveniente implementar una interfaz gráfica de usuario

7.2. Trabajos futuros

(GUI). Esto premitirá a los profesionales sanitarios introducir textos o diagnósticos de forma directa y obtener informes resumidos de una manera más intuitiva, sin tener conocimientos específicos de programación.

A. Apéndice

A.1. Ejemplo de diagnóstico de un paciente

Name: ____ Unit No: ____

Admission Date: ____ Discharge Date: ____

Date of Birth: ____ Sex: M

Service: CARDIOTHORACIC

Allergies:

Penicillins

Attending: ____.

Chief Complaint:

chronic cough

Major Surgical or Invasive Procedure:

Right VATS wedge resection x 3

History of Present Illness:

Mr ___ is a ___ with symptoms
of chronic wheeze, cough and SOB x ___ yrs. He was diagnosed with
asthma, but he also has had interstitial findings on CT since
___ with minimal variability. His IgE is elevated, has
eosinophilia even on prednisone, ___ wnl, so it is unclear
what his exact diagnosis is. He is sent for eval for possible
VATS biopsy.

Past Medical History:

PAST MEDICAL HISTORY:

Alcohol dependence 303.90

Anxiety states 300.00

Hypercholesterolemia 272.0

IMPOTENCE DUE TO ERECTILE DYSFUNCTION 607.84

Colonic polyp 211.3

A. Apéndice

Esophageal reflux 530.81
Lung disease, chronic obstructive 496
DRUG DEPENDENCE - IN REMISSION(aka DRUG) 304.73
SLEEP APNEA, UNSPEC 780.57
OPTIC PAPILLITIS 377.31
DECLINED - COPD (NOT DX, FOR PROB LIST ONLY) PRGEN16
Obesity 278.00
Squamous cell skin cancer 173.92
Eosinophilia 288.3
asthma 493.90
History of tobacco use V15.82
Multiple allergies V15.09
Osteoma ear, s/p surg
acute sinusitis treated w/ levaquin ___
recent fractured ribs left 7,8,9 from fall
? I&D right leg as child

Social History:

Family History:

Mother: died cancer in ___ (pt not sure of type)

Father

___: brother w/ leukemia

Offspring

Other

Physical Exam:

BP: 138/79. Heart Rate: 87. Weight: 215.8. Height: 71. BMI:

30.1.

Temperature: 97.6. Resp. Rate: 16. Pain Score: 0. 02

Saturation%:

97.

GENERAL

[x] WN/WD [x] NAD [x] AAO [x] abnormal findings: very ruddy complexion

HEENT

[x] NC/AT [x] EOMI [x] PERRLA [x] Anicteric

[x] OP/NP mucosa normal [x] Tongue midline

[x] Palate symmetric [x] Neck supple/NT/without mass

[x] Trachea midline [x] Thyroid nl size/contour

[x] Abnormal findings: poor dentition

RESPIRATORY

[x] CTA/P [x] Excursion normal [x] No fremitus

A.1. Ejemplo de diagnóstico de un paciente

[x] No egophony [x] No spine/CVAT

[] Abnormal findings:

CARDIOVASCULAR

[x] RRR [x] No m/r/g [x] No JVD [x] PMI nl [x] No edema

[x] Peripheral pulses nl [x] No abd/carotid bruit

[] Abnormal findings:

GI

[x] Soft [x] NT [x] ND [x] No mass/HSM [x] No hernia

[] Abnormal findings:

GU [x] Deferred

[] Nl genitalia [] Nl pelvic/testicular exam [] Nl DRE

[] Abnormal findings:

NEURO

[x] Strength intact/symmetric [x] Sensation intact/ symmetric

[x] Reflexes nl [x] No facial asymmetry [x] Cognition intact

[x] Cranial nerves intact [] Abnormal findings:

MS

[x] No clubbing [x] No cyanosis [x] No edema [x] Gait nl

[x] No tenderness [x] Tone/align/ROM nl [x] Palpation nl

[x] Nails nl [] Abnormal findings:

LYMPH NODES

[x] Cervical nl [x] Supraclavicular nl [x] Axillary nl

[x] Inguinal nl [] Abnormal findings:

SKIN

[x] No rashes/lesions/ulcers

[x] No induration/nodules/tightening [] Abnormal findings:

PSYCHIATRIC

[x] Nl judgment/insight [x] Nl memory [] Nl mood/affect

[x] Abnormal findings: irritable

Pertinent Results:

___ CXR :

Right chest tube projects at the apex, and there is no pneumothorax.

A. Apéndice

Brief Hospital Course:

Mr. ___ was admitted to the hospital and taken to the Operating Room where he underwent a right VATS with multiple wedge resections. He tolerated the procedure well and returned to the PACU in stable condition. He maintained stable hemodynamics and his pain was well controlled.

Following transfer to the Surgical floor he required the addition of Ibuprofen for better pain control. He had better effect and was able to use his incentive spirometer effectively. His morning CXR had a tiny apical PTX although he had no respiratory symptoms. His chest tube had no air leak and was removed once his chest tube drainage diminished. His post-pull film had a stable apical PTX. He was tolerating a regular diet and walking independently. He was discharged home without any respiratory symptoms.

Medications on Admission:

The Preadmission Medication list is accurate and complete.

1. PredniSONE 5 mg PO DAILY
2. Tiotropium Bromide 1 CAP IH DAILY
3. Montelukast Sodium 10 mg PO DAILY
4. Albuterol Inhaler 2 PUFF IH Q6H:PRN wheezing
5. Fluticasone-Salmeterol Diskus (500/50) 1 INH IH BID

Discharge Medications:

1. Albuterol Inhaler 2 PUFF IH Q6H:PRN wheezing
 2. Fluticasone-Salmeterol Diskus (500/50) 1 INH IH BID
 3. Montelukast Sodium 10 mg PO DAILY
 4. PredniSONE 5 mg PO DAILY
 5. Tiotropium Bromide 1 CAP IH DAILY
 6. Acetaminophen 650 mg PO Q6H
 7. Docusate Sodium 100 mg PO BID
 8. Ibuprofen 600 mg PO Q6H:PRN pain
take with food
- RX *ibuprofen 600 mg 1 tablet(s) by mouth four times a day Disp #*100 Tablet Refills:*1
9. OxycoDONE (Immediate Release) ___ mg PO Q4H:PRN pain
- RX *oxycodone 5 mg ___ tablet(s) by mouth every four (4) hours Disp #*60 Tablet Refills:*0

Discharge Disposition:

Home

Discharge Diagnosis:

Interstitial lung disease

A.1. Ejemplo de diagnóstico de un paciente

Discharge Condition:

Mental Status: Clear and coherent.

Level of Consciousness: Alert and interactive.

Activity Status: Ambulatory - Independent.

Discharge Instructions:

- * You were admitted to the hospital for lung surgery and you've recovered well. You are now ready for discharge.
- * Continue to use your incentive spirometer 10 times an hour while awake.
- * Check your incisions daily and report any increased redness or drainage. Cover the area with a gauze pad if it is draining.
- * Your chest tube dressing may be removed in 48 hours. If it starts to drain, cover it with a clean dry dressing and change it as needed to keep site clean and dry.
- * You will continue to need pain medication once you are home but you can wean it over a few weeks as the discomfort resolves. Make sure that you have regular bowel movements while on narcotic pain medications as they are constipating which can cause more problems. Use a stool softener or gentle laxative to stay regular.
- * No driving while taking narcotic pain medication.
- * Take Tylenol ___ mg every 6 hours in between your narcotic. If your doctor allows you may also take Ibuprofen to help relieve the pain.
- * Continue to stay well hydrated and eat well to heal your incisions
- * Shower daily. Wash incision with mild soap & water, rinse, pat dry
- * No tub bathing, swimming or hot tubs until incision healed
- * No lotions or creams to incision site
- * Walk ___ times a day and gradually increase your activity as you can tolerate.

Call Dr. ___ if you experience:

-Fever > 101 or chills

-Increased shortness of breath, chest pain or any other symptoms that concern you.

Followup Instructions:

A. Apéndice

A.2. Ejemplo de la primera experimentación: resumen de consulta clínica

En esta sección se recoge un ejemplo del primer experimento de generación de textos clínicos. En él se indica el input (prompt) y la salida devuelta por el modelo para la prueba con rol específico y sin rol. Para ello se le ha pedido al modelo, que basado en el contexto (texto indicado como entrada), devuelva un resumen del diagnóstico del paciente. En este ejemplo, el modelo utilizado ha sido Gemma a temperatura 1.0.

Input (prompt) al modelo

Based on

Name: ___ Unit No: ___

Admission Date: ___

Discharge Date: ___

Date of Birth: ___

Sex: M

Service: CARDIOTHORACIC

Allergies:

diltiazem / ether

Attending: ___.

Chief Complaint:

Left Lower Lung Nodule

Major Surgical or Invasive Procedure:

Left thoracoscopy with wedge resection and mediastinal node sampling.

History of Present Illness:

The patient is a ___ year old man who is referred to thoracic surgery by Dr. ___ for evaluation of a small pulmonary nodule that was detected on a CT done for evaluation of an aortic aneurysm. Mr. ___ has been followed by Dr. ___ ___ COPD and sleep apnea. In ___, he underwent stent grafting of an abdominal aneurysm. A CT from three months earlier showed a 100 left lower lobe nodule which, on the more recent scan (___ shows it to have increased in size to 13mm. The patient has mild exertional dyspnea and occasionally uses oxygen at home

Past Medical History:

PAST MEDICAL HISTORY: Includes high cholesterol, hypertension,

A.2. Ejemplo de la primera experimentación: resumen de consulta clínica

atrial fibrillation, heart valve disease. Positive PPD, COPD.

PAST SURGICAL HISTORY: Includes an ORIF to the left hip in

---.
He has additionally suffered a tibial plateau fracture on the left and left medial malleolar fracture on the left.

Social History:

Family History:

Includes high cholesterol, hypertension and renal failure.

Physical Exam:

VS: Temp: 99.3 (Tm 99.3), BP: 111/70 (93-156/54-97), HR: 130 (75-130), RR: 18, O₂ sat: 94% (89-94), O₂ delivery: RA (1L-2L)

GEN: Well appearing, no acute distress

HEENT: NCAT, EOMI, sclera anicteric

CV: HDS

PULM: No signs of respiratory distress. Incisions c/d/I. Prior CT

dressing clean and dry.

EXT: Warm, well-perfused

NEURO: A&Ox3, no focal neurologic deficits

Pertinent Results:

___ 04:25AM BLOOD WBC-11.4* RBC-4.72 Hgb-14.2 Hct-44.4

MCV-94 MCH-30.1 MCHC-32.0 RDW-13.4 RDWSD-45.9 Plt ___

___ 05:14AM BLOOD WBC-10.4* RBC-4.88 Hgb-14.9 Hct-45.6

MCV-93 MCH-30.5 MCHC-32.7 RDW-13.3 RDWSD-45.4 Plt ___

___ 04:25AM BLOOD Plt ___

___ 05:14AM BLOOD Plt ___

___ 04:25AM BLOOD Glucose-116* UreaN-22* Creat-1.1 Na-135

K-4.5 Cl-96 HC03-28 AnGap-11

___ 05:14AM BLOOD Glucose-141* UreaN-21* Creat-1.0 Na-135

K-4.5 Cl-98 HC03-28 AnGap-9*

___ 04:25AM BLOOD Calcium-9.1 Phos-3.1 Mg-2.2

___ 05:14AM BLOOD Calcium-9.1 Phos-3.6 Mg-1.____ male patient with an incident of being off left lower lung nodule and PET scan showing an FDG avid nodule who is admitted to the hospital for surgical management. On ___

he underwent thoracoscopy and left lower lung wedge resection.

Please report to the operative note for further detail. After an uncomplicated procedure he was sent to the floor and started on a regular diet. On postop day 1 his chest x-ray showed no pneumothorax and his chest tube had no early for which the chest tube was removed. The post pull x-ray showed no pneumothorax or

A. Apéndice

changes from previous chest x-ray. During that day he had intermittent episodes of A. fib SVTs that were well controlled with his home medication of metoprolol. Today he is tolerating a regular diet without nausea or emesis, he denies shortness of breath, he reports pain is well controlled, has vital signs within normal limits, his Xarelto has been restarted, and is ambulating without difficulty. He is being discharged in good condition.

Medications on Admission:

The Preadmission Medication list is accurate and complete.

1. BuPROPion (Sustained Release) 150 mg PO QAM
2. Tiotropium Bromide 1 CAP IH DAILY
3. Metoprolol Succinate XL 100 mg PO DAILY
4. Alendronate Sodium 70 mg PO QSUN
5. Verapamil SR 120 mg PO Q24H
6. Rivaroxaban 20 mg PO DAILY
7. Atorvastatin 40 mg PO QPM
8. Multivitamins 1 TAB PO DAILY
9. TraMADol 50 mg PO BID:PRN Pain - Moderate
10. Fluticasone-Salmeterol Diskus (500/50) 1 INH IH BID
11. albuterol sulfate 90 mcg/actuation inhalation TID:PRN wheezing

Discharge Medications:

1. Acetaminophen 1000 mg PO Q6H
RX *acetaminophen 500 mg 2 tablet(s) by mouth every six (6) hours Disp #*100 Tablet Refills:*0
2. Docusate Sodium 100 mg PO BID
RX *docusate sodium 100 mg 1 capsule(s) by mouth twice a day Disp #*60 Capsule Refills:*2
3. OxyCODONE (Immediate Release) ___ mg PO Q4H:PRN Pain - Moderate
RX *oxycodone 5 mg ___ tablet(s) by mouth Q ___ hrs Disp #*40 Tablet Refills:*0
4. albuterol sulfate 90 mcg/actuation inhalation TID:PRN wheezing
5. Alendronate Sodium 70 mg PO QSUN
6. Atorvastatin 40 mg PO QPM
7. BuPROPion (Sustained Release) 150 mg PO QAM
8. Fluticasone-Salmeterol Diskus (500/50) 1 INH IH BID
9. Metoprolol Succinate XL 100 mg PO DAILY
10. Multivitamins 1 TAB PO DAILY
11. Rivaroxaban 20 mg PO DAILY
12. Tiotropium Bromide 1 CAP IH DAILY
13. TraMADol 50 mg PO BID:PRN Pain - Moderate
use this if its more effective than Oxycodone

A.2. Ejemplo de la primera experimentación: resumen de consulta clínica

14. Verapamil SR 120 mg PO Q24H

Discharge Disposition:

Home

Discharge Diagnosis:

Left lower lobe lung nodule.

Discharge Condition:

Mental Status: Clear and coherent.

Level of Consciousness: Alert and interactive.

Activity Status: Ambulatory - Independent.

Discharge Instructions:

* You were admitted to the hospital for lung surgery and you've recovered well. You are now ready for discharge.

* Continue to use your incentive spirometer 10 times an hour while awake.

* Check your incisions daily and report any increased redness or drainage. Cover the area with a gauze pad if it is draining.

* Your chest tube dressing may be removed in 48 hours. If it starts to drain, cover it with a clean dry dressing and change it as needed to keep site clean and dry. You have some stitches in place that will need to be removed when you see Dr. ___ on ____.

* You may need pain medication once you are home but you can wean it over the next week as the discomfort resolves. Make sure that you have regular bowel movements while on narcotic pain medications as they are constipating which can cause more problems. Use a stool softener or gentle laxative to stay regular.

* No driving while taking narcotic pain medication.

* Take Tylenol on a standing basis to avoid more opioid use.

* Continue to stay well hydrated and eat well to heal your incisions

* No heavy lifting > 10 lbs for 4 weeks.

A. Apéndice

- * Shower daily. Wash incision with mild soap & water, rinse, pat dry
 - * No tub bathing, swimming or hot tubs until incision healed
 - * No lotions or creams to incision site

- * Walk ___ times a day and gradually increase your activity as you can tolerate.

Call Dr. ___ ___ if you experience:

- Fevers > 101 or chills
- Increased shortness of breath, chest pain or any other symptoms that concern you.

** If pathology specimens were sent at the time of surgery, the reports will be reviewed with you in detail at your follow up appointment. This will give both you and your doctor time to understand the pathology, its implications and discuss options going forward.**

Followup Instructions:

- . Give me a summary of the diagnostic for the patient.

Output del modelo

■ Sin rol específico:

Summary of Diagnostic Findings:

This patient is a 60-year-old male admitted for a left lower lung nodule.

Key Diagnoses:

* **Left lower lobe lung nodule:** The patient was referred for evaluation of a small pulmonary nodule detected on a CT scan for an aortic aneurysm. The nodule has increased in size and is causing mild exertional dyspnea.

* **Possible malignancy:** The nodule is FDG avid on PET scan, indicating it may

A.3. Ejemplo de la segunda experimentación: prueba de ventana de contexto

be cancerous. Further pathology reports are needed to confirm this.

****Other Relevant Information:****

* **History:** The patient has a history of COPD, sleep apnea, atrial fibrillation, heart valve disease, hypertension, and high cholesterol. He also has

■ Con rol específico:

This patient is a 60-year-old male admitted for a left lower lung nodule.

****Key Diagnoses:****

* **Left lower lobe lung nodule:** The patient's initial presentation was a small pulmonary nodule detected on a CT scan for an aortic aneurysm. The nodule has grown in size and is now 13mm.
* **Possible malignancy:**

****** The nodule is FDG avid on a PET scan, indicating it may be cancerous. The patient is scheduled for a follow-up with Dr. [Name] to discuss pathology results and treatment options.

****Additional Information:****

* **History:** The patient has a history of COPD, sleep apnea, hypertension, atrial fibrillation, heart valve disease

A.3. Ejemplo de la segunda experimentación: prueba de ventana de contexto

En esta sección se recoge un ejemplo del segundo experimento de generación de textos clínicos. En él se indica el input (prompt) y la salida devuelta por el modelo sin rol. Para ello se le ha pedido al modelo, que basado en el contexto (texto indicado como entrada), devuelva en formato json el diagnóstico del paciente y la derivación, es decir, si se fue al hospital, ICU o a su casa. En este ejemplo, el modelo utilizado ha sido Gemma a temperatura 1.0.

Input (prompt) del modelo

Based on Sex: F

Service: UROLOGY

A. Apéndice

Allergies:

Patient recorded as having No Known Allergies to Drugs

Attending: ___.

Chief Complaint:

renal mass

Major Surgical or Invasive Procedure:

right laparoscopic radical nephrectomy- Dr. ___, Dr.

___ ___

History of Present Illness:

___ y/o healthy female with incidental finding of right renal mass suspicious for RCC following MRI on ___.

Past Medical History:

PMH: nonspecific right axis deviation

PSH- cesarean section

ALL-NKDA

Social History:

Family History:

no history of RCC

Pertinent Results:

___ 07:15AM BLOOD WBC-7.6 RBC-3.82* Hgb-11.9* Hct-33.8*

MCV-89 MCH-31.2 MCHC-35.2* RDW-12.8 Plt ___

___ 07:15AM BLOOD Glucose-150* UreaN-10 Creat-0.9 Na-138

K-3.8 Cl-104 HC03-27 AnGap-11

Brief Hospital Course:

Patient was admitted to Urology after undergoing laparoscopic right radical nephrectomy. No concerning intraoperative events occurred; please see dictated operative note for details. The patient received perioperative antibiotic prophylaxis. The patient was transferred to the floor from the PACU in stable condition. On POD0, pain was well controlled on PCA, hydrated for urine output >30cc/hour, provided with pneumoboots and incentive spirometry for prophylaxis, and ambulated once. On POD1, foley was removed without difficulty, basic metabolic panel and complete blood count were checked, pain control was transitioned from PCA to oral analgesics, diet was advanced to a

A.3. Ejemplo de la segunda experimentación: prueba de ventana de contexto

clears/toast and crackers diet. On POD2, diet was advanced as tolerated. The remainder of the hospital course was relatively unremarkable. The patient was discharged in stable condition, eating well, ambulating independently, voiding without difficulty, and with pain control on oral analgesics. On exam, incision was clean, dry, and intact, with no evidence of hematoma collection or infection. The patient was given explicit instructions to follow-up in clinic with ___ in 3 weeks.

Medications on Admission:

none

Discharge Medications:

1. Hydrocodone-Acetaminophen ___ mg Tablet Sig: ___ Tablets PO Q6H (every 6 hours) as needed for break through pain only (score >4) .

Disp:*60 Tablet(s)* Refills:*0*

2. Docusate Sodium 100 mg Capsule Sig: One (1) Capsule PO BID (2 times a day).

Disp:*60 Capsule(s)* Refills:*2*

Discharge Disposition:

Home

Discharge Diagnosis:

renal cell carcinoma

Discharge Condition:

stable

Discharge Instructions:

-You may shower but do not bathe, swim or immerse your incision.

-Do not eat constipating foods for ___ weeks, drink plenty of fluids

-Do not lift anything heavier than a phone book (10 pounds) or drive until you are seen by your Urologist in follow-up

-Tylenol should be used as your first line pain medication. If your pain is not well controlled on Tylenol you have been prescribed a narcotic pain medication. Use in place of Tylenol. Do not exceed 4 gms of Tylenol in total daily

A. Apéndice

-Do not drive or drink alcohol while taking narcotics

-Resume all of your home medications, except hold NSAID (aspirin, advil, motrin, ibuprofen) until you see your urologist in follow-up

-If you have fevers > 101.5 F, vomiting, or increased redness, swelling, or discharge from your incision, call your doctor or go to the nearest ER

-Call Dr. ___ to set up follow-up appointment and if you have any urological questions. ___

Followup Instructions:

. In json format, write in a line: diagnostic:, and what is the diagnostic for the patient. In other 1

Output del modelo

```
```json
{
 "diagnostic": "Renal Cell Carcinoma",
 "solution": "None"
}
```
```

Bibliografía

- [BLM⁺24] Sally L Baxter, Christopher A Longhurst, Marlene Millen, Amy M Sitapati, y Ming Tai-Seale. Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned. *JAMIA Open*, 7(2):ooaeo28, 04 2024.
- [CK23] Dillon Chrimes y Chanhee Kim. Comparison of mimic-iii and mimic-iv for big data analytics of health informatics. En *2023 IEEE International Conference on Big Data (BigData)*, páginas 6128–6130, 2023.
- [CSCW20] Zhihong Chen, Yan Song, Tsung-Hui Chang, y Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.
- [GDJ⁺24] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhota, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthi, Sheng Shen, Shengye Wan, Shruti Bhowmik, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov,

Bibliografía

Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcau-te, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojaezeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie

- Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wencheng Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaoqian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, y Zhiyu Ma. The llama 3 herd of models, 2024.
- [GGC⁺22] Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, y Rahmatollah Beheshti. An extensive data processing pipeline for mimic-iv, 2022.
- [HIP] <https://cloud.google.com/security/compliance/hipaa-compliance?hl=es>.
- [HPBP⁺24] Emman Haider, Daniel Perez-Becker, Thomas Portet, Piyush Madan, Amit Garg, Atabak Ashfaq, David Majercak, Wen Wen, Dongwoo Kim, Ziyi Yang, Jianwen Zhang, Hiteshi Sharma, Blake Bullwinkel, Martin Pouliot, Amanda Minnich, Shiven Chawla, Solianna Herrera, Shahed Warreth, Maggie Engler, Gary Lopez, Nina Chikanov, Raja Sekhar Rao Dheekonda, Bolor-Erdene Jagdagdorj, Roman Lutz, Richard Lundein, Tori Westerhoff, Pete Bryan, Christian Seifert, Ram Shankar Siva Kumar, Andrew Berkley, y Alex Kessler. Phi-3 safety post-training: Aligning language models with a "break-fixcycle, 2024.
- [iSa22] iSanidad. Casi el 70 % de la carga de trabajo en la consulta se traduce en labores administrativas, 2022. [Recurso online, accedido el 3 de junio de 2025].
- [JBS⁺23] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.
- [JSM⁺23] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, y William El Sayed. Mistral 7b, 2023.
- [JSR⁺24] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
- [LSX⁺22] Renqian Luo, Lliai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, y Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), September 2022.
- [PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, y Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, páginas 311–318, 2002.
- [QY⁺25] Qwen, : An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, y Zihan Qiu. Qwen2.5 technical report, 2025.
- [Rep23] Mari Reponen. Predicting prescribed medications from the mimic-iv event and measurement data. 2023.

Bibliografía

- [RNS⁺18] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [Sho77] Edward H Shortliffe. Mycin: A knowledge-based computer program applied to infectious diseases. En *Proceedings of the Annual Symposium on Computer Application in Medical Care*, página 66, 1977.
- [TLI⁺23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambo, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [Top19] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.
- [TRP⁺24] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kociský, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeannie Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, y Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024.
- [VSP⁺23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, y Illia Polosukhin. Attention is all you need, 2023.

Bibliografía

- [VVVUB⁺23] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Małgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. Clinical text summarization: adapting large language models can outperform human experts. *Research square*, páginas rs–3, 2023.
- [YLY⁺25] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yingger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, y Zihan Qiu. Qwen3 technical report, 2025.
- [ZKW⁺20] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, y Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.