UNIVERSIDAD POLITÉCNICA SALESIANA
SEDE QUITO
CARRERA DE BIOMEDICINA

GENERACIÓN DE INFORMES RADIOLÓGICOS MEDIANTE EL RECONOCIMIENTO DE VOZ.

Trabajo de titulación previo a la obtención del Título de:
Ingeniero Biomédico

AUTOR: Ilan Patricio Benalcázar Cando

TUTOR: Ph.D. Fabián Rodrigo Narváez Espinoza

Quito-Ecuador

2025

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Yo, Ilan Patricio Benalcázar Cando con documento de identificación No. 1004451595 expreso mi voluntad y por medio del presente documento cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que soy autor del Proyecto Técnico: GENERACIÓN DE INFORMES RADIOLÓGICOS MEDIANTE EL RECONOCIMIENTO DE VOZ, el cual ha sido desarrollado para optar por el título de: Ingeniero Biomédico, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribo este documento en el momento que hago la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Quito, 4 de noviembre del año 2025.

Atentamente,

Ilan Patricio Benalcázar Cando
1004451595

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Yo, Ilan Patricio Benalcázar Cando con documento de identificación N° 1004451595 manifiesto que:

Soy el autor y responsable del presente trabajo; y, autorizo a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Quito, 4 de noviembre del año 2025.

Atentamente,

Ilan Patricio Benalcázar Cando
1004451595

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Fabián Rodrigo Narváez Espinoza con documento de identificación N° 0103674677, docente de la Universidad Politécnica Salesiana declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: GENERACIÓN DE INFORMES RADIOLÓGICOS MEDIANTE EL RECONOCIMIENTO DE VOZ, realizado por Ilan Patricio Benalcázar Cando con documento de identificación N° 1004451595, obteniendo como resultado final el trabajo de titulación bajo la opción Proyecto Técnico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Quito, 05 de noviembre del año 2025.

Atentamente,

Ph.D. Fabián Rodrigo Narváez Espinoza
0103674677

LEMA

"No importa el problema,
mientras vivas tienes una oportunidad"

Agradecimientos

Con mi más profundo estima, agradecimiento y benquerencia a mis padres Maila Milisenjoset Cando Rendon y Patricio Ernesto Benalcázar León por darme la oportunidad y financiar mis estudios, así como de los materiales necesarios para el desarrollo de este proyecto. También mi más profunda gratitud a mi director de tesis el Dr. Fabián Rodrigo Narváez Espinoza por su guía siendo un pilar fundamental para el desarrollo de esta tesis. Finalmente mi más sincera gratitud a mi prima Ariana Daniela Rivadeneira Cando por su retroalimentación y criticas constructivas con respecto al proyecto.

Resumen

La generación de informes radiológicos es un proceso crítico que, en muchos servicios, aún depende del dictado y la redacción manual o de soluciones ASR genéricas poco adaptadas al español clínico. Estos enfoques presentan desventajas recurrentes: latencias elevadas en la elaboración del reporte, errores ortográficos y de acentuación, confusiones terminológicas (acrónimos y abreviaturas), inconsistencias de estilo y, cuando se recurre a servicios en la nube, restricciones de privacidad y dependencia de conectividad. La variabilidad dialectal, el ruido ambiental y la falta de estandarización agravan el problema, incrementando la carga de posedición del radiólogo y afectando la eficiencia del flujo de trabajo. Por ello, se requiere una solución local, modular y especializada que reduzca errores, mantenga el registro radiológico y respete la confidencialidad de los datos.

En este proyecto se implementó un sistema automático y on-premise para la generación de informes radiológicos en español a partir de dictado, integrado por cinco etapas: (i) adquisición de audio y preprocesamiento acústico (normalización y remuestreo) para reducir la variabilidad de entrada; (ii) transcripción automática con Whisper-small (Transformer multilingüe) configurado para español y con marcas temporales por palabra para asegurar trazabilidad; (iii) revisión léxica asistida por un modelo biomédico tipo RoBERTa, destinado a normalizar términos clínicos, diacríticos, acrónimos y unidades; (iv) adecuación estilística mediante Gemma-2B-IT ajustado finamente con LoRA/QLoRA (cuantización de 4 bits para eficiencia), con el fin de mantener la estructura y el tono propios del informe radiológico sin introducir hallazgos no dictados; y (v) postprocesamiento y entrega del informe con una interfaz gráfica que expone versiones intermedias (borrador, correcciones y final) y facilita la pos–edición por el especialista.

La evaluación se realizó sobre un conjunto de 100 informes con pares prompt–target (dictado con errores vs. versión corregida por experto). Frente a la configuración base sin especialización, el sistema propuesto mostró mejoras sustanciales en métricas automáticas: incrementos de hasta +112 % en BLEU y +77 % en ROUGE-2, además de aumentos consistentes en ROUGE-1 y ROUGE-L, lo que se traduce en menor necesidad de correcciones y mejor preservación de la estructura narrativa (hallazgo → localización → características). La arquitectura local demostró ser desplegable en CPU/GPU modestas, disminuyendo la dependencia de servicios externos y fortaleciendo la confidencialidad. En conjunto, la propuesta confirma la factibilidad técnica y la utilidad aplicada del enfoque para acelerar la elaboración de informes, elevar la consistencia terminológica y reducir la carga de pos–edición, sentando una base robusta para extensiones futuras por subespecialidad y para la incorporación de métricas clínicas específicas.

Palabras clave: reconocimiento de voz, informes radiológicos, modelos de lenguaje, fine-tuning, Whisper, Gemma, NLP clínico.

Abstract

The generation of radiology reports is a critical process that, in many departments, still relies on manual dictation and transcription or on generic ASR solutions poorly adapted to clinical Spanish. These approaches have recurrent drawbacks: high latency in report preparation, spelling and accentuation errors, terminological inconsistencies (acronyms and abbreviations), style variability, and, when cloud-based services are used, privacy concerns and dependency on connectivity. Dialectal variability, background noise, and lack of standardization further aggravate the problem, increasing the radiologist's post-editing workload and reducing workflow efficiency. Therefore, a local, modular, and specialized solution is required—one that minimizes errors, preserves radiology records, and ensures data confidentiality.

In this project, we implemented an on-premise automatic system for generating radiology reports in Spanish from dictated input, composed of five stages: (i) audio acquisition and acoustic preprocessing (normalization and resampling) to reduce input variability; (ii) automatic transcription using Whisper-small (a multilingual Transformer) configured for Spanish, with word-level timestamps to ensure traceability; (iii) lexical review assisted by a biomedical RoBERTa-based model aimed at normalizing clinical terms, diacritics, acronyms, and measurement units; (iv) stylistic adaptation using Gemma-2B-IT fine-tuned with LoRA/QLoRA (4-bit quantization for efficiency) to preserve the typical structure and tone of radiology reports without introducing undictated findings; and (v) postprocessing and report delivery through a graphical interface that displays intermediate versions (draft, corrected, and final) and facilitates expert post-editing.

Evaluation was conducted on a set of 100 reports consisting of prompt–target pairs (error-prone dictation vs. expert-corrected version). Compared to the baseline configuration without specialization, the proposed system showed substantial improvements in automatic metrics: up to +112 % increase in BLEU and +77 % in ROUGE-2, along with consistent gains in ROUGE-1 and ROUGE-L. These results indicate reduced correction effort and better preservation of the narrative structure (finding → location → characteristics). The local architecture proved deployable on modest CPU/GPU setups, reducing reliance on external services and enhancing confidentiality. Overall, the proposal demonstrates both the technical feasibility and applied usefulness of the approach to accelerate report generation, improve terminological consistency, and reduce post-editing workload—laying a solid foundation for future extensions by subspecialty and the integration of clinical performance metrics.

Keywords: speech recognition, radiology reports, language models, fine-tuning, Whisper, Gemma, clinical NLP.

Contenido

Lista de Figuras

Lista de Tablas

Lista de símbolos

Símbolos con letras latinas

Símbolo	Término	Definición / referencia
S	Sustituciones	Nº de palabras sustituidas. Ecuación 2-1
D	Eliminaciones	Nº de palabras eliminadas. Ecuación 2-1
I	Inserciones	Nº de palabras insertadas. Ecuación 2-1
N	Nº de palabras de referencia	Denominador en la ecuación2-1
BP	*Brevity Penalty*	Factor de penalización por longitud 2-2
p_n	Precisión modificada de n-gramas	Razón de coincidencias de n-grama. Ecuación 2-2
w_n	Peso de cada n-grama	En BLEU, $w_n = 1/N$ 2-2
LCS	*Longest Common Subsequence*	Longitud de la subsecuencia común. Ecuación **??**

Subíndices

Subíndice	Significado
n	Tamaño del n-grama (BLEU, ROUGE)
i	Índice de palabra o de n-grama en una secuencia

Abreviaturas

Abreviatura	**Término**
ASR	Reconocimiento Automático de Habla
BLEU	Subestudio de evaluación bilingüe
CAS	Puntuación de precisión clínica
CNN	Red Neuronal Convolucional
DICOM	Imágenes y Comunicaciones Digitales en Medicina Digital
DNN	Red Neuronal Profunda
EHR	Historia clínica electrónica
LLM	Modelos Largos de Lenguaje
NLP	Procesamiento de Lenguaje Natural
PACS	Sistema de Archivo y Comunicación de Imágenes
RIS	Sistema de Información Radiológica
RNN	Red Neuronal Recurrente
ROUGE	Subestudio orientado al recuerdo para la evaluación de Gisting
WER	Tasa de error de palabra

1 Introducción

El informe radiológico es un documento que comunica los hallazgos de un estudio de imagen al médico tratante, siendo fundamental para el diagnóstico y manejo terapéutico del paciente. En la práctica clínica convencional, la generación manual de estos informes es un proceso laborioso y propenso a errores que impacta negativamente tanto en la calidad de la atención al paciente como en la eficiencia del servicio de radiología. Cada radiólogo dedica entre 15 y 20 minutos a dictar y redactar manualmente cada informe tras la exploración visual de las imágenes, lo que representa una carga significativa de tiempo y un cuello de botella en el flujo de trabajo de los servicios de radiología (Alqahtani et al., 2024).

Este enfoque produce una alta probabilidad de errores humanos, como duplicaciones de hallazgos, omisiones de datos relevantes o inconsistencias en la terminología, que resultan en informes poco claros y diagnósticos imprecisos. Tales errores no solo retrasan la entrega de los diagnósticos, sino que también aumentan el riesgo de decisiones clínicas equivocadas, especialmente en escenarios de alta complejidad o cuando los profesionales cuentan con menor experiencia (Bitterman et al., 2021). Además, la acumulación de estos errores humanos provoca demoras en la atención al paciente e incrementa la carga laboral de los radiólogos. La falta de estandarización en la estructura y el lenguaje de los informes agrava el problema, dificultando su integración con los sistemas de historia clínica electrónica y comprometiendo la trazabilidad y reutilización de la información para estudios posteriores o investigación clínica (Chew & Ngiam, 2025).

Con el propósito de reducir estos fallos, las primeras aproximaciones para automatizar la generación de informes radiológicos se basaron en sistemas de reconocimiento de voz convencionales acoplados a procesadores de lenguaje natural (NLP) basados en reglas. Sin embargo, estos sistemas tienen problemas para reconocer con precisión el vocabulario especializado y para adaptarse a las particularidades del habla de cada radiólogo. Su dependencia a diccionarios predefinidos limitaba severamente su capacidad para manejar la variabilidad del dictado clínico. En la práctica esto se refleja en la transcripción incorrecta de términos médicos que suenen parecidos como "tosz "todos"podían confundirse, y acrónimos como "TEP"(Tromboembolismo Pulmonar) se transcribían literalmente o se interpretaban erróneamente. Asimismo, la variación en el acento, el tono de voz o la presencia de ruido ambiental generaban errores que comprometían la calidad del informe, especialmente en medidas numéricas críticas donde "4.5 cm"podía transcribirse como "45 cm". Adicionalmente el flujo de trabajo aumentaba, ya que por los problemas descritos con anterioridad se necesitaba de una corrección

manual constante del informe transcrito por estos softwares lo que anulaba gran parte de las ganancias de eficiencia esperadas. Estas limitaciones dificultaban que el sistema se adaptara adecuadamente a la terminología propia del área en donde se desarrollaba y a las particularidades del modo de hablar de cada radiólogo (Czum, 2020).

Con el auge de los modelos de aprendizaje automático, las técnicas de procesamiento de lenguaje natural (NLP) comenzaron a incorporar el aprendizaje profundo para mejorar la precisión en la transcripción y comprensión del dictado clínico. Estos sistemas utilizan .embeddings", que son representaciones vectoriales de palabras que capturan su significado semántico en una matriz, permitiendo que términos médicos con significados similares tengan representaciones cercanas. En aplicaciones clínicas especializadas como la oncología radioterápica, se observó que las redes neuronales recurrentes combinadas con embeddings médicos mejoraban significativamente la identificación de entidades clínicas y relaciones complejas dentro del texto. Sin embargo, estos enfoques presentaban limitaciones importantes, como la necesidad de extensos volúmenes de datos etiquetados para el entrenamiento de embeddings específicos del dominio médico (Bitterman et al., 2021).

Posteriormente, la introducción de arquitecturas de redes neuronales conocidas como Transformer revolucionó el campo del análisis de imágenes médicas y la generación de texto asociado. Estas arquitecturas utilizan mecanismos de auto-atención que permiten evaluar simultáneamente la importancia de diferentes regiones dentro de una imagen o secciones de texto. Las investigaciones han demostrado que estos enfoques tienen superioridad en tareas de segmentación y detección de hallazgos médicos en comparación con arquitecturas convolucionales tradicionales (Azad et al., 2024). Al mismo tiempo, se desarrollaron generadores automáticos de informes radiológicos que combinan técnicas de mejora de contraste en imágenes con modelos Transformer capaces de aprender detalles y relaciones entre características visuales y descripciones textuales (Tsaniya et al., 2024). No obstante, los modelos Transformer presentan problemas como: requisitos computacionales elevados o la necesidad de grandes conjuntos de datos de entrenamiento (Azad et al., 2024).

El desarrollo de arquitecturas híbridas que combinan CNNs (Redes Neuronales Convolucionales) para extracción de características visuales con Transformers para generación de lenguaje. En este contexto, se crearon frameworks, los cuales son entornos de desarrollo que proporcionan herramientas, librerías y estructuras estandarizadas, diseñados específicamente para el diagnóstico consciente de la anatomía. Estos frameworks incorporan conocimiento estructural en el proceso de generación de reportes torácicos, lo que mejora sustancialmente la coherencia clínica de los informes producidos (Li et al., 2024). Simultáneamente, surgieron arquitecturas que fusionan operaciones convolucionales con mecanismos de atención multi-cabeza para abordar el desequilibrio en los datos y enriquecer la contextualización de hallazgos. Estas arquitecturas híbridas pese a sus ventajas enfrentan complejidades técnicas notables, particularmente en la integración eficiente entre mecanismos de atención y operaciones convolucionales, así

como en el balance del entrenamiento entre los componentes de visión computacional y procesamiento de lenguaje natural. Esto implica que cuando estos sistemas no logran una sincronización adecuada entre los flujos visuales y textuales, los informes generados pueden presentar incongruencias clínicas, descripciones imprecisas de estructuras anatómicas o interpretaciones erróneas de lesiones específicas. (Alqahtani et al., 2024).

Para la documentación automática, los estudios sistemáticos de interacciones clínicas registradas en historiales electrónicos han demostrado que los métodos basados en redes neuronales profundas proporcionan una cobertura más completa de notas médicas y reducen significativamente los errores en comparación con sistemas basados en reglas (Falcetta et al., 2023). Recientemente, la exploración de grandes modelos de lenguaje (LLMs) para reportes estructurados en radiología ha evidenciado su potencial para generar secciones completas de hallazgos y conclusiones con un alto grado de coherencia y adaptabilidad (Busch et al., 2024). Sin embargo, los LLMs también presentan limitaciones importantes, incluyendo la tendencia a generar contenido incorrecto o .ªlucinaciones", sensibilidad a variaciones en las instrucciones y desafíos en el manejo consistente de terminología médica especializada (Busch et al., 2024).

En el campo de la integración de reconocimiento automático del habla (ASR) con generación de lenguaje, se ha demostrado la viabilidad de utilizar modelos avanzados para la elaboración de resúmenes de alta complejidad, como informes de alta hospitalaria y reportes operatorios, mostrando fortalezas en fluidez textual pero también limitaciones en el control de terminología especializada (Dubinski et al., 2024). Los sistemas integrados enfrentan dificultades que incluyen la propagación de errores de transcripción a través de las fases de procesamiento, manejo inconsistente de terminología médica especializada y problemas para mantener la coherencia clínica (Dubinski et al., 2024).

Importancia y alcance

El presente trabajo describe un sistema automático de generación de informes radiológicos ejecutable en Visual Studio Code con herramientas gratuitas y de código abierto. En primer lugar, la voz del radiólogo almacenada en un archivo .wav en español se transcribe con un modelo de inteligencia artificial conocido como Whisper (modelo openai/whisper-small), que además proporciona una puntuación de confianza por palabra. Los tokens (unidad básica y elemental del modelo de lenguaje) con confianza baja se enmascaran en la secuencia resultante. Seguido, dichas máscaras se corrigen de forma contextual con el modelo PlanTL-GOB-ES/lm-biomedical-clinical-es en modo fill-mask, lo que permite que este BERT (Behavioral Emergency Response Team) clínico especializado en español sustituya cada palabra dudosa por la alternativa más probable según su conocimiento lingüístico biomédico. Finalmente, la transcripción es revisada y reescrita por el modelo Gemma 2-it fine-tuneado para garantizar coherencia clínica y formato estructurado.

El sistema ofrece una alternativa económica y accesible para la generación automatiza-

da de informes radiológicos en comparación con soluciones comerciales que suelen ser costosas y requieren infraestructura especializada. La utilización de software de código abierto y modelos pre-entrenados disponibles públicamente permite un manejo más sencillo, facilitando la adopción de estas tecnologías en entornos clínicos.

1.1. Objetivos

1.1.1. Objetivo General

- Desarrollar un sistema de reconocimiento de voz a texto con terminología especializada para generar informes radiológicos.

1.1.2. Objetivo Específicos

- Desarrollar un prototipo funcional que transcriba y corrija dictados radiológicos con un modelo largo de lenguaje fine tuneado.
- Implementar un modelo de NLP(Procesamiento de Lenguaje Natural), con el contexto radiológico.
- Evaluar la precisión de la transcripción de terminología radiológica.

2 Procesamiento Natural del Lenguaje: Marco Teórico

En la última década, la creación de informes en el área de radiología ha avanzado gracias al procesamiento natural del lenguaje (NLP por sus siglas en ingles) que han evolucionado desde sistemas rígidos basados en reglas gramaticales a las redes neuronales capaces de aprender e integrar diferentes tipos de información. Los primeros sistemas de NLP basados en reglas eran programados con reglas gramaticales estrictas lo que hacia que se tengan que actualizar a cada momento, debido a que el lenguaje humano está en constante cambio combinado con los acentos de cada persona, así como también su falta de adaptabilidad gracias a las reglas gramaticales por las cuales fueron programados, provocaban que cuando estos modelos eran llevados al campo del léxico médico fallaran al reconocer sinónimos o abreviaturas (Fanni et al., 2024; Luo & Chong, 2020). Adicionalmente, al ser sistemas que dependían de multiples actualizaciones hacían que los costos de su mantenimiento se incrementara (Czum, 2020).

Con la advenimiento de las redes neuronales recurrentes (RNN por sus siglas en ingles) con mecanismos de atención y su implementación en el campo de la radiología oncológica (Bitterman et al., 2021), se ha demostrado que estos modelos pueden reducir errores hasta un 20 % y que podían procesar las largas secuencias de terminología usada en el diagnóstico, pero la naturaleza de estos modelos al ser secuenciales tenían el problema de no poder recibir datos diferentes o multimodales, por ejemplo una imagen y texto al mismo tiempo, lo cual hacia que su utilidad en contextos más complejos, como el seguimiento de enfermedades crónicas decreciera y siga sin ser un modelo verdaderamente útil (Zhang et al., 2024).

En años recientes desde la llegada de Chat GPT se empezó a observar un incremento en el uso de una arquitectura de redes neuronales denominada Transformers, la cual resulta ser útil gracias a sus mecanismos de atención global que hacen que estos modelos no tiendan a perder la información, sino también a contextualizarla de manera eficiente en secuencias largas de información que se pueden aplicar tanto como para texto o imágenes al mismo tiempo. En investigaciones recientes, se ha demostrado la capacidad de los Transformers para analizar imágenes médicas, relacionando patrones presentes en estudios de resonancia magnética con la terminología empleada en los informes radiológicos, lo que evidencia su potencial multimodal (Azad et al., 2024). De igual manera, trabajos previos han mostrado el uso de modelos Transformers multimodales en tareas de segmentación de tumores pancreáticos y en la redacción automática de

descripciones radiológicas, integrando de manera coherente información visual y textual (He et al., 2023).

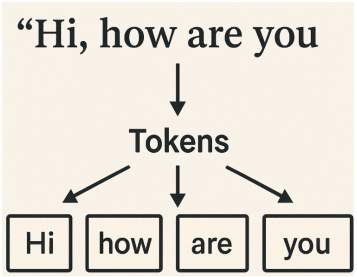
Por otro lado, el aprendizaje profundo ha demostrado con la incorporación de conceptos médicos anatómicos puede mejorar la precisión en los informes quirúrgicos, debido a que las redes neuronales, logran establecer patrones que permiten corregir términos de manera más precisa (Groot et al., 2021). Asimismo, se han desarrollado modelos que combinan reglas lingüísticas con arquitecturas de Transformers más ligeras para optimizar la comprensión, es decir aquellas expresiones lingüísticas que permiten identificar la secuencia y el momento en que ocurren los eventos descritos en los textos médicos (Olex & McInnes, 2021). Por otro lado, la inclusión de conceptos anatómicos ha resultado esencial para priorizar los hallazgos en órganos relevantes, fortaleciendo así la interpretación clínica mediante el uso de estos conceptos (Li et al., 2024). Finalmente, se han propuesto modelos de lenguaje de gran tamaño de tipo multimodal, capaces de extraer y correlacionar automáticamente información proveniente tanto de imágenes médicas como de textos, demostrando el potencial de éstas arquitecturas para la integración de datos médicos complejos (Zhang et al., 2024).

2.1. Fundamento del Procesamiento Natural del Lenguaje

El funcionamiento del Procesamiento del Lenguaje Natural (PLN) se fundamenta en convertir secuencias lingüísticas humanas en representaciones numéricas comprensibles para modelos de aprendizaje automático. A diferencia de los sistemas tradicionales basados en reglas, los actuales modelos de PLN permiten representar el significado semántico y sintáctico de las expresiones lingüísticas de manera continua y distribuida, facilitando tareas como la clasificación de texto, la generación automática de resúmenes o la estructuración de conversaciones clínicas.

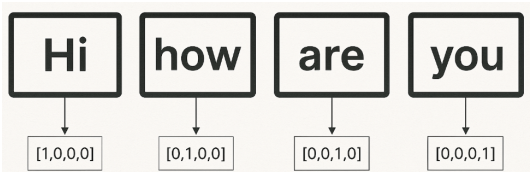
El primer paso en cualquier sistema moderno de PLN es la tokenización, es decir, la división del texto en unidades discretas llamadas *tokens*. Dependiendo del enfoque, un token puede corresponder a una palabra, una subpalabra o un carácter como se ilustra en la Figura 1. Los esquemas tradicionales de tokenización por palabras demostraron ser insuficientes frente a lenguas con una estructura compleja o ante palabras raras o términos especializados. Por ello, en modelos modernos como BERT o GPT (Devlin et al., 2019; Yenduri et al., 2023), se recurre a la tokenización por subpalabras, mediante algoritmos como Byte Pair Encoding (BPE) o WordPiece (Kudo, 2018; Sennrich et al., 2016), que permiten representar cualquier palabra como una combinación de fragmentos frecuentes aprendidos del corpus de entrenamiento.

Figura 1. *Ilustración de ejemplo de tokenización con la frase "Hi, how are you".*



Una vez dividido el texto en tokens, a cada uno se le asigna un número único que lo identifica dentro del vocabulario del modelo el cual se utiliza para indexar una matriz de *embeddings*. Esta matriz es un diccionario matemático donde cada token tiene asociado un vector de embedding; dicho vector es una lista de números en un espacio de alta dimensionalidad que captura el significado semántico y las relaciones contextuales del token, talk como se ilustra en la Figura 2. Los *word embeddings* permiten capturar similitudes semánticas y relaciones sintácticas entre palabras. Por ejemplo, modelos como Word2Vec y GloVe (Mikolov et al., 2013; Pennington et al., 2014), demostraron que relaciones como rey - hombre + mujer = reina emergen de forma natural en estos espacios.

Figura 2. *Ilustración de ejemplo de embeding con la frase "Hi, how are you".*



Si bien los embeddings estáticos capturan ciertas propiedades generales de las palabras, tienen la limitación de asignar un único vector a cada palabra, sin considerar su contexto. Para superar esto, modelos como BERT introdujeron representaciones *contextuales*, donde el embedding de una palabra depende de las palabras que la rodean (Devlin et al., 2019; Peters et al., 2018). En estos modelos, cada token pasa a través de una arquitectura tipo Transformer que utiliza mecanismos de autoatención para integrar información contextual y genera el texto.

En aplicaciones médicas, estos mecanismos permiten a los modelos adaptarse a la variabilidad del lenguaje utilizado en notas clínicas, reportes radiológicos o conversaciones paciente-médico. Los LLMs basados en embeddings contextuales pueden reconocer sinónimos, corregir ambigüedades y establecer relaciones entre términos anatómicos y hallazgos, mejorando la precisión de tareas como la generación automática de informes.

2.2. Modelos Largos de Lenguaje (LLMs)

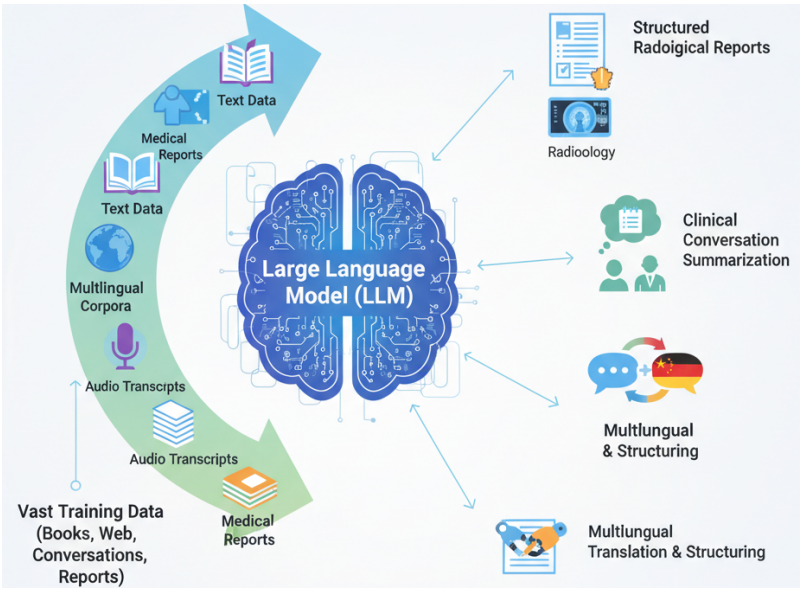
Los Large Language Models (LLMs) son modelos de lenguaje basados en arquitecturas transformer, entrenados en grandes volúmenes de datos textuales para comprender, generar y manipular lenguaje humano. Estos son capaces de producir texto de manera coherente y lógica, y de seguir instrucciones emitidas por el operador Busch et al., 2025.

Entre los modelos más representativos se encuentran GPT-3.5, GPT-4 y, más recientemente, Gemma 2 (Team et al., 2024), una familia de modelos abiertos desarrollada por Google DeepMind. Gemma 2 se distingue por su eficiencia y escalabilidad, ofreciendo tamaños reducidos (2B–27B parámetros) que permiten su implementación en entornos médicos con recursos computacionales limitados. Su arquitectura optimizada conserva capacidades avanzadas de comprensión semántica y generación contextual, lo que la convierte en una alternativa para aplicaciones clínicas y de investigación biomédica.

Estos modelos se han utilizado para automatizar la creación de informes radiológicos estructurados (SR), lo que mejora la eficiencia y la precisión en la documentación clínica (Delbrouck et al., 2025). Pueden transformar texto libre en informes estructurados, reduciendo errores y aumentando la adherencia a las guías clínicas. Por ejemplo, mediante GPT-4 se logró una precisión del 100 % en la conversión de informes de tomografía computarizada (TC) y resonancia magnética (RM) a formatos estructurados sin pérdida de información (Busch et al., 2025). Además, los LLMs han mostrado capacidades multilingües, permitiendo la traducción y estructuración de informes en diferentes idiomas, como francés, italiano, chino y alemán, lo que facilita la estandarización de los informes radiológicos (Busch et al., 2024; Jelassi et al., 2024).

Los LLMs también se han aplicado en la transcripción de conversaciones clínicas entre médicos y pacientes, lo que ayuda a identificar riesgos de salud y apoyar la toma de decisiones. Modelos de lenguaje como BERT (Wang & Zhang, 2024), T5 y BioGPT han sido utilizados para extraer información relevante de diálogos médicos y generar resúmenes precisos. Por ejemplo, la técnica Cluster2Sent (Krishna et al., 2021), que emplea el modelo de lenguaje T5 como base generativa, ha demostrado ser efectiva en la creación de resúmenes a partir de transcripciones de consultas médicas (Krishna et al., 2021).

Figura 3. *Ilustración general de un modelo largo de lenguaje y sus aplicaciones para organizar y resumir información.*



2.2.1. Ajuste fino

El ajuste fino (fine-tuning) es una estrategia especialmente útil para adaptar modelos pre-entrenados en tareas específicas o algún dominio en particular con la ayuda de conjuntos de datos más reducidos. Este proceso permite mejorar el rendimiento de un modelo ligero al enseñarle a partir de predicciones generadas por un modelo más potente, una técnica que se asemeja a la destilación del conocimiento en el aprendizaje profundo tradicional (Zhang et al., 2024). Una de las ventajas del ajuste fino es que elimina la necesidad de entrenar modelos desde cero, porque el conocimiento general del lenguaje ha sido previamente aprendido. Posteriormente, se pueden añadir capas específicas para la tarea y ajustar el modelo para otras diversas aplicaciones (Moezzi et al., 2023).

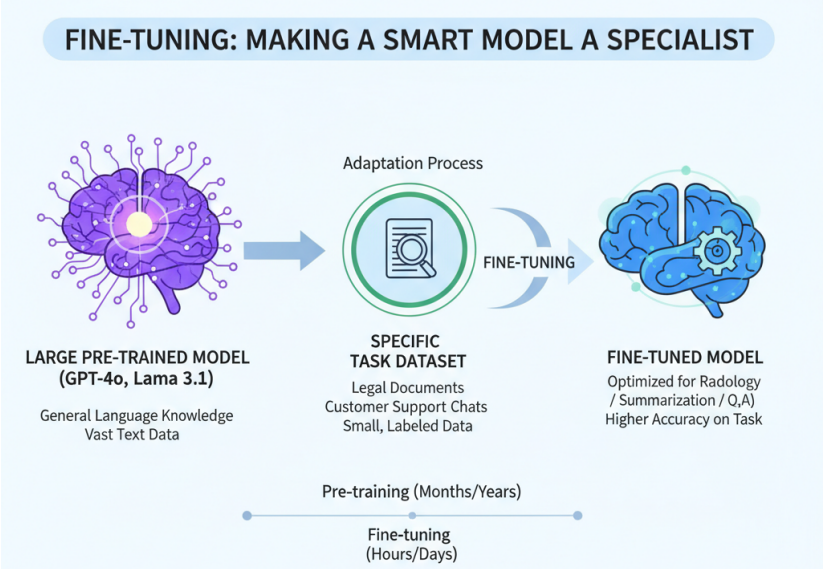
En el contexto de los informes radiológicos, elajuste fino ha demostrado ser una estrategia que supera limitaciones como el tamaño de los modelos y la escasez de datos. Se ha observado su potencial para mejorar LLMs ligeros, como Llama 3.1-8B (Grattafiori et al., 2024), utilizando conjuntos de datos con etiquetas sintéticas o débilmente etiquetadas. Esta técnica ha permitido que los modelos ajustados superen incluso a sus modelos bases o más grandes, lo que resalta la capacidad intrínseca del modelo para aprender de datos de calidad variable.

Además, el ajuste fino se ha aplicado en diversas tareas dentro del dominio radiológico. En un estudio previo (Guo et al., 2023), se utilizó para la clasificación de enfermedades pulmonares de opción múltiple en informes radiológicos, donde el modelo predecía enfermedades de una lista predeterminada . Para ello, se empleó un etiquetador para extraer etiquetas y construir instrucciones para el ajuste fino. Así mismo, el ajuste

fino se ha utilizado en la detección de enfermedades pulmonares de formato abierto, requiriendo que el LLM extraiga hallazgos anómalos de informes radiológicos, utilizando etiquetas sintéticas generadas por modelos como GPT-4o (Busch et al., 2025). La combinación de conjuntos de instrucciones derivados de diferentes conjuntos de datos permite un ajuste fino conjunto de un único LLM para optimizar su rendimiento en múltiples tareas (Guo et al., 2023; Wei et al., 2024).

Además, el ajuste fino ha sido empleado en la transformación de informes radiológicos de texto libre a formatos estructurados, usando modelos como T5 y Scifive (una adaptación pre-entrenada de T5 específica para el dominio) que han sido ajustados con una pequeña cantidad de datos anotados de informes radiológicos para extraer entidades y relaciones. Este enfoque basado en transformers ha superado a métodos basados en redes neuronales artificiales (ANN) y las redes neuronales convolucionales (CNN) en la generación de informes estructurados interpretables (Moezzi et al., 2023).

Figura 4. *Ilustración del proceso de fine-tuning: un modelo general se adapta a tareas específicas para mejorar su desempeño.*



2.2.2. Arquitectura de sistemas ASR médicos

El reconocimiento automático de habla (ASR) en entornos clínicos implica tres componentes principales:

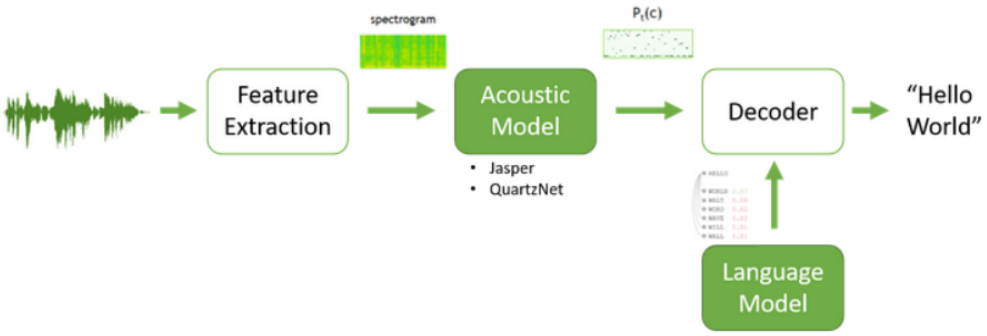
- **Modelo acústico**: Transforma señales de audio en unidades fonéticas mediante redes neuronales profundas (DNN).

- **Modelo de lenguaje**: Asigna probabilidades a secuencias de términos médicos usando atención global de Transformers .
- **Decodificador**: Optimiza la transcripción mediante algoritmos como lo puede ser beam search.

En la Figura 5 se ilustra la arquitectura general de un sistema ASR de sus siglas en inglés Automatic Speech Recognition, donde el proceso se inicia con la extracción de características del audio inicial, que son representadas como espectrogramas. Estas características se procesan por el modelo acústico, el cual calcula la probabilidad de secuencias fonéticas, después el decodificador con ayuda de un modelo de lenguaje convierten esas secuencias en texto coherente. Esta arquitectura puede adaptarse a contextos clínicos mediante el entrenamiento con terminología médica especializada (Jorg et al., 2023; NVIDIA, 2020).

Si bien la Figura 5 describe el enfoque "por componentes", en la práctica clínica también se emplean modelos *end-to-end* que unifican el modelo acústico, el de lenguaje y el decodificador bajo un único Transformer. Un ejemplo destacado es Whisper (Radford et al., 2022): un encoder–decoder Transformer que recibe espectrogramas log-Mel y genera texto de manera autorregresiva. Para convertir probabilidades en texto, Whisper usa estrategias de decodificación como *greedy* seleccionar en cada paso el token con mayor probabilidad, lo que es muy rápido y estable pero puede caer en soluciones no optimas al ignorar alternativas futuras y *beam search* mantener varias hipótesis en paralelo (el "haz") y ampliar solo las más prometedoras según su probabilidad acumulada, lo que explora más combinaciones y suele producir frases más coherentes, a cambio de mayor coste computacional. En entornos médicos, su salida puede complementarse con adaptación al dominio (p. ej., ajuste fino o *prompts*/vocabularios especializados) y post-procesadores léxicos para asegurar la correcta normalización de terminología sensible.

Figura 5. *Ilustración de la arquitectura general de un sistema ASR con modelos acústico, de lenguaje y decodificador*



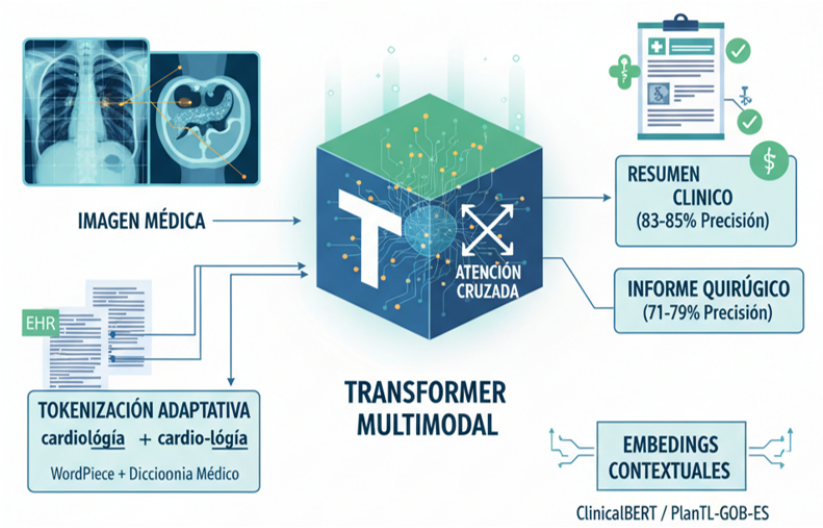
Note. *Tomado de NVIDIA (2020).*

2.2.3. Modelos Transformers para análisis de imágenes médicas y procesamiento de lenguaje especializado

Los modelos Transformers enfocados en imágenes, también conocidos como Vision Transformers, han superado el desempeño de las redes neuronales convolucionales (CNN, por sus siglas en inglés), al lograr una segmentación más precisa y una detección más efectiva de patrones sutiles en imágenes médicas, como radiografías torácicas y resonancias magnéticas de páncreas (Guo et al., 2023; Ozsahin et al., 2025). Esta superioridad se debe a la capacidad de los Transformers para analizar las relaciones entre los píxeles de una imagen y para abordar tareas multimodales que integran texto e imagen, lo cual ha permitido su aplicación en la generación automatizada de historias clínicas electrónicas (EHR, por sus siglas en inglés) (Azad et al., 2024; Wang & Zhang, 2024).

En el área del Procesamiento del Lenguaje Natural (NLP) especializado, las palabras se dividen en sílabas o letras que contienen la información recibida de manera específica según el caso (conocidos como token). Este proceso combina métodos como WordPiece (separar en sílabas o letras) con vocabularios provenientes de diccionarios médicos para permitir que el modelo maneje terminología especializada y abreviaturas (Czum, 2020). Mediante esta técnica, se logró una reducción del 18 % en los errores de procesamiento. Por otra parte, modelos como ClinicalBERT y PlanTL-GOB-ES (Carrino et al., 2022), generan vectorizaciones de la información (embeddings) contextuales debido a que fueron pre-entrenados con terminología biomédica. Esto mejora la codificación de términos complejos en español, en el caso del mmodelo PlanTL-GOB-ES, y en inglés, en el caso del modelo ClinicalBERT y sus variantes. Para las tareas multimodales, se emplea una arquitectura con atención cruzada, la cual permite al modelo establecer relaciones entre dos fuentes o secuencias de datos distintas, tal como se ilustra en la Figura 8. De esta forma, se integran características visuales y textuales, lo que posibilita alcanzar altos niveles de concordancia en resúmenes clínicos y en informes quirúrgicos (Alqahtani et al., 2024; Zhang et al., 2024).

Figura 6. *Esquema de un transformer multimodal que integra imagen y texto para apoyar el análisis clínico.*



2.2.4. Arquitecturas híbridas

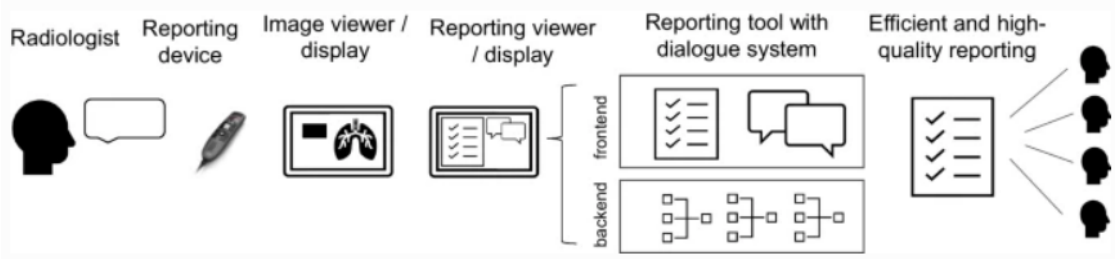
En los últimos años, las arquitecturas híbridas han sido un pilar importante a la hora de realizar informes radiológicos de manera automática, para tal efecto estas arquitecturas combinan las CNN y Transformers, respectivamente (Alqahtani et al., 2024). Esto es debido a que las CNN son buenas para el análisis de imágenes, y que pueden extraer representaciones jerárquicas, mientras que los Transformers usan mecanismo de atención que evitan la falta de memoria, lo cual se logra capturando relaciones semánticas complejas entre diferentes tipos de datos cuando se maneja grandes cantidades de los mismos (He et al., 2023). Por lo tanto, la combinación de arquitecturas Transformers con las CNN supera a sistemas que solo trabajen con una sola arquitectura de manera independiente. esto se ha demostrado que esta combinación reporta buen desempeño en tareas como la clasificación, segmentación y resumen en imágenes médicas.

Este enfoque multimodal, desarrollado mediante sistemas híbridos, ha sido validado en investigaciones donde se emplearon estos modelos para comparar características clínicas, como los antecedentes médicos de los pacientes con rasgos radiológicos extraídos de radiografías torácicas (Guo et al., 2023; Li et al., 2024). En dicho estudio, se aplicaron técnicas de codificación visual-clínica y wisdom learning, las cuales integran el conocimiento experto humano con algoritmos de aprendizaje automático, lo que permite contextualizar los datos médicos dentro de marcos diagnósticos validados (Iqbal et al., 2024).

Por muy efectivos que resultan estos sistemas, no están exentos de problemas como el costo computacional, que se tiene requiere al entrenar estos modelos y ejecutar, lo que se refleja por la necesidad de una gran cantidad de memoria gráfica. Sin embargo este problema puede resolverse con el uso de plataformas en la nube. De igual manera,

estos sistemas pueden hacer uso de una conexión directa con el sistema de archivado y comunicación de imágenes (PACS por sus siglas en ingles) y el sistema de información radiológica (RIS por sus siglas en ingles)(Jorg et al., 2024). Una de las principales mejoras de estos sistemas es el trabajo organizado de manera modular para identificar un problema, actualizarse o adaptarse al contexto y los diferentes flujos de trabajo en ciertas áreas de aplicación (Lin & Kuo, 2025).

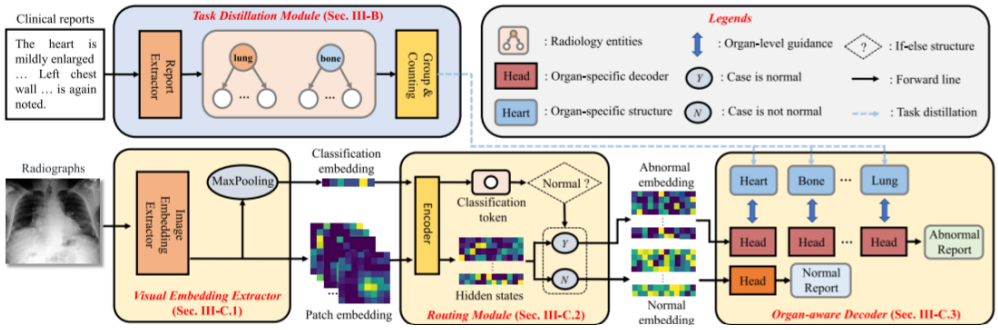
Figura 7. *Arquitectura modular del flujo de trabajo ASR-NLP*



Note. La etapa de ASR transcribe el dictado en bruto (1), el módulo NLP normaliza términos y estructura los hallazgos (2), y el generador multimodal integra imágenes DICOM con texto (3). Tomado de Jorg et al. (2023).

Las aplicaciones en distintas subespecialidades evidencian la versatilidad de este enfoque. En nefrorradiología, se ha demostrado que un modelo puede diferenciar quistes simples de masas sólidas mediante el uso de mapas de atención sobre regiones renales (Sheikhy et al., 2025). En cardiología pediátrica, otros investigadores optimizaron embeddings con el fin de identificar anomalías congénitas en los flujos sanguíneos (Taylor, 2022). Actualmente, se han implementado modelos LLMs especializados en oftalmología, los cuales alcanzaron un 94 % de concordancia con oftalmólogos en la interpretación de retinografías diabéticas (Yang et al., 2024). Estos avances resaltan la importancia del fine-tuning específico por dominio y la disponibilidad de corpus clínicos anotados por expertos para cada aplicación médica.

Figura 8. *Ilustración de la arquitectura multimodal para generación de informes*



Note. Tomado de (Li et al., 2024).

2.3. Medidas de evaluación de desempeño

El desempeño de los modelos de lenguaje natural han sido evaluados de manera cuantitativa y de forma objetiva, lo que nos determina la calidad y eficacia de cada modelo. En términos generales, las medidas de desempeño comparan las salidas generadas por los modelos frente a los términos verdaderos que debe reportar. Es por eso que es importante establecer de esta forma el verdadero desempeño de los modelos utilizados. Para tal efecto se describen las siguientes medidas de desempeño:

2.3.1. Word Error Rate (WER)

El *Word Error Rate* (WER) cuantifica la discrepancia entre la transcripción generada por un sistema de reconocimiento automático de habla (ASR) y una referencia humana mediante la fórmula:

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100\,\%, \tag{2-1}$$

donde S son sustituciones, D eliminaciones, I inserciones y N el total de palabras en la referencia. En radiología, valores de WER cercanos al 17 % se han logrado al adaptar Whisper al francés clínico, mientras que bases de datos especializadas reducen aún más el error al introducir terminología propia del dominio (Alharbi et al., 2021; Jelassi et al., 2024). Sin embargo, WER penaliza alteraciones que no siempre afectan la interpretación diagnóstica; por ello, en la evaluación de informes finales se recurre a métricas que miden coincidencias léxicas y estructurales, como BLEU y ROUGE. Las métricas Sustituto de evaluación bilingüe (BLEU) y el Sustituto orientado al recuerdo para la evaluación de Gisting (ROUGE) se emplean para comparar la salida de un modelo con informes redactados por especialistas (Casey et al., 2021).

2.3.2. Sustituto de evaluación bilingüe (BLEU).

BLEU calcula la precisión de n-gramas de la hipótesis (H) con respecto a la referencia (R), penalizando cadenas demasiado cortas mediante un *brevity penalty* (BP):

$$\text{BLEU-}N = BP\ \exp\Big(\sum_{n=1}^{N} w_n\ \log p_n\Big), \tag{2-2}$$

con $w_n = 1/N$ y p_n la precisión modificada de n-gramas. Modelos ajustados con Transformers han alcanzado BLEU ≈ 0.74 en ecografía abdominopélvica, frente a valores de 0.10 en enfoques basados en atención visual sobre radiografías de tórax (Babar et al., 2021; Moezzi et al., 2023). Un sistema reciente entrenado para generaciones estructuradas (CheXpert-Plus) eleva el índice a 14.8 % en impresiones clínicas (Delbrouck et al., 2025).

2.3.3. Sustituto orientado al recuerdo para la evaluación de Gisting (ROUGE).

ROUGE-N mide el *recall* de n-gramas de la referencia capturados por el candidato:

$$\text{ROUGE-}N = \frac{\sum_{\text{ref}} \sum_{g_n \in \text{ref}} \text{Count}_{\text{match}}(g_n)}{\sum_{\text{ref}} \sum_{g_n \in \text{ref}} \text{Count}(g_n)} \times 100\,\%, \tag{2-3}$$

mientras que, ROUGE-L se define como la media armónica entre la precisión (P) y la sensibilidad (R) calculados a partir de las subsecuencias comunes entre el candidato y la referencia, y no se basa directamente en la longitud de la subsecuencia común más larga:

$$\text{ROUGE-}L = \frac{2PR}{P+R} \tag{2-4}$$

En tareas de conversión de reportes libres a formatos estructurados se han informado valores de ROUGE-L de 0.53, mientras que en conjuntos de tórax las variantes palabra-a-palabra rondan 0.29 y los sistemas orientados a resúmenes estructurados superan el 28 % (Babar et al., 2021; Delbrouck et al., 2025; Moezzi et al., 2023). Pese a su popularidad, BLEU y ROUGE pueden sobreestimar la calidad si la coincidencia léxica no implica equivalencia diagnóstica, lo que ha dado lugar a métricas específicas (p. ej., F1-RadGraph o DCS) para validar el contenido clínico de los informes.

Al combinar WER para la fase de transcripción con BLEU y ROUGE para la etapa de generación de informes, se obtiene un marco integral de evaluación técnica. No obstante, la fidelidad semántica sigue exigiendo métricas adicionales orientadas al dominio radiológico, garantizando así la utilidad clínica de los sistemas automatizados (Casey et al., 2021; Jorg et al., 2024).

3 Sistema de generación de informes radiológicos mediante reconocimiento de voz

El sistema propuesto en este trabajo es una aplicación de escritorio para la transcripción y corrección asistida de dictados radiológicos en español. Su arquitectura es modular y con procesos que transforman una grabación de voz en un informe final de texto coherente y terminológicamente consistente. Para ello el sistema integra 5 etapas que son: I) La adquisición de audio. II) Etapa de normalización y preprocesamiento. III) Motor de transcripción (ASR). IV) Una revisión semántica con conocimiento biomédico. V) La generación del informe final.

Esta separación por etapas permite sustituir o mejorar componentes sin afectar al resto del flujo, favorece la trazabilidad mediante registros de cada fase y brinda al usuario una interfaz gráfica que muestra tanto las transcripciones intermedias como el resultado consolidado. La Figura 9 ilustra el flujo de trabajo propuesto.

Figura 9. *Ilustración de la arquitectura modular del sistema propuesto para transcripción y corrección radiológica*

3.1. Arquitectura del sistema propuesto

A continuación se describen cada una de las etapas o módulos que conforman la arquitectura general del sistema propuesto.

3.1.1. Módulo 1: Adquisición de audio

Este módulo se encarga de la captura del audio dictado y almacenamiento como un archivo para el sistema. Este modulo fue desarrollado de manera que la captura puede realizarse de dos maneras: grabación en tiempo real desde el micrófono o carga de un archivo existente. En ambos casos, el propósito es conservar una copia fiel del audio original, acompañada de metadatos mínimos.

Grabación en tiempo real

Al iniciar la grabación se inicia un flujo continuo de datos del micrófono (stream). La señal de voz analógica se convierte en valores numéricos mediante el convertidor analógico–digital (ADC) de la tarjeta de audio. Para este propósito se implementó la libreria conocida como, `sounddevice` (Hans et al., 2025), que hace que el dispositivo de audio entregue el sonido en pequeños paquetes (*buffers*) de muestras a intervalos muy cortos; cada vez que llega un paquete, se ejecuta automáticamente una función (*callback*) que lo recoge y lo copia en la memoria de la aplicación. Además para realizar estas tareas, también se implementó el tipo de dato `float32` mediante la utilización de la librería `NumPy`(Developers, 2025), la misma que permite cambiar la frecuencia a 16 kHz y 1 canal. Esta se caracteriza por su baja latencia, control explícito de parámetros (tasa de muestreo y canales) y amplia compatibilidad con drivers, lo cual es esencial para dictado médico continuo. Por otro lado, `sounddevice` recibe del sistema pequeños paquetes de audio y, cada vez que llega uno, llama a la función (*callback*); esta función únicamente añade el paquete a una cola en la memoria RAM. Como se procesan paquetes cortos y se devuelve el control de inmediato, el sistema propuesto continúa ejecutándose sin bloquearse.

Estos paquetes de muestras (*frames*) se van guardando temporalmente en memoria. Al detener la grabación, todos los paquetes son unidos en una sola secuencia. De igual manera se usa la librería `NumPy`, para la conversión masiva (vectorizada) de datos tipo `float32` a PCM de 16 bits para que sea rápida y segura, es decir: escala el audio normalizado al rango entero propio del formato, redondea, recorta cualquier valor que se salga del rango permitido y convierte el arreglo al tipo entero correspondiente. Esta librería ha sido implementada debido a su alto rendimiento y a que aplica el escalado correcto evitando desbordes/*clipping*. Asímismo, se ha implementado la librería `wave` (Foundation, 2025b), para el manejo de archivos de tipo `.wav` bien formado (cabecera + datos PCM) sin compresión. El archivo se guarda en `recorded_audio/` con un nombre basado en fecha y hora, y permite el registro de metadatos: ruta, duración y tasa de

muestreo reportada por el dispositivo. Finalmente, este módulo incluye un sistema para registro de eventos e incidencias (por ejemplo, micrófono no disponible).

Carga de archivos

Una vez que se selecciona un audio existente (`.wav`, `.mp3`, `.m4a`, `.ogg`, `.flac`), el módulo verifica que el archivo sea legible, obtiene su tamaño y extrae metadatos básicos del contenedor (formato, canales y tasa de muestreo declarados). En este trabajo se implementó un proceso para leer esos metadatos sin decodificar todo el audio el cual sigue los siguientes pasos: (1) para `.wav` se implementó el lector estándar `wave.open`, el cual es una función de la librería `wave` que hace que podamos leer la cabecera de formato de archivo de intercambio de recursos o RIFF (por sus siglas en ingles) (Corporation, 2025) y, a partir de ella, obtener el codificador (`codec`), el número de canales (`nchannels`), la tasa de frames (`framerate`) y el `tamaño de muestra`; se implementó dicho lector debido a que basta con leer la cabecera para conocer los parámetros con exactitud y no requiere dependencias mucha carga computacional. RIFF es un contenedor que organiza el archivo en "chunks" o bloques; en `.wav` los chunks más relevantes son `fmt` (describe códec, número de canales, tasa de muestreo y profundidad de bits) y `data` (contiene las muestras). (2) para `.mp3`/`.m4a`/`.ogg`/`.flac` se implementó `ffprobe` el cual es una herramienta del framework FFmpeg y sirve para extraer información técnica, esta herramienta es llamada por la libreria `subprocess` (Foundation, 2025a) con salida a un archivo con un formato de texto ligero y legible por humanos para almacenar e intercambiar datos de manera estructurada JSON, que hace que el propio contenedor entregue su códec, número de canales y tasa de muestreo de forma uniforme; se lo usó debido a su compatibilidad entre múltiples formatos sin depender de bibliotecas distintas para cada contenedor.

Funciones claves del módulo

Este modulo tiene asignado algunas funciones claves para asegurar la calidad de la información capturada, las mismas que se describen a continuación: Inicializar la fuente de audio (micrófono o archivo); capturar o leer la señal sin pérdidas; acumular bloques cortos y consecutivos de audio en memoria mientras se graba; normalizar el resultado a un archivo `.wav` cuando corresponda; generar un nombre único y conservar la ruta; extraer y almacenar metadatos mínimos; y registrar en la bitácora los eventos e incidencias de la adquisición.

Caso de Uso

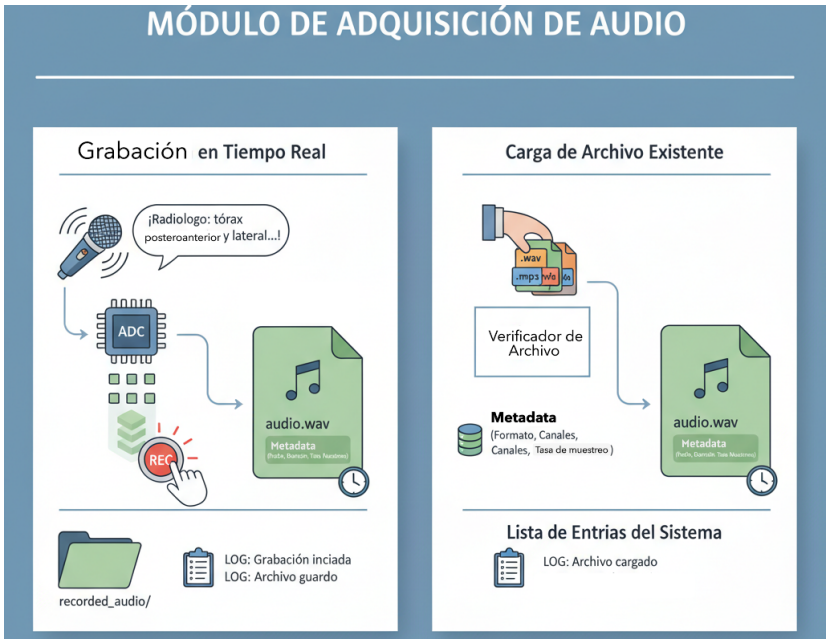
Sí el radiólogo describe oralmente: "tórax posteroanterior y lateral muestra un parénquima pulmonar".

1. Se inicia la grabación. El micrófono entrega la señal al ADC, que la convierte en números; el módulo forma frames y va acumulando en memoria mientras se dicta la frase.
2. Al detener, los frames se concadenan y se escriben en un archivo `.wav` dentro de

`recorded_audio/` (por ejemplo, `2025-08-02_233500.wav`). Se guardan la ruta, la duración y la tasa de muestreo reportada.

3. El resultado de la etapa es un archivo con el dictado completo y sus metadatos, listo para ser utilizado por el resto del sistema.

Figura 10. *Ilustración del Módulo de adquisición de audio.*



3.1.2. Módulo 2: Normalización y preprocesamiento de los datos de ingreso

En este módulo se estandariza la señal acústica proveniente de la adquisición para que toda entrada cumpla un perfil interno único el cual es: tener un solo canal de reproducción del audio o también llamado "mono", una frecuencia de 16kHz, PCM lineal de 16 bits, es un método que convierte una señal de sonido continua en una secuencia de valores numéricos que representan su amplitud en distintos instantes, es decir, una señal discreta) y nivel de referencia estable. Las operaciones se aplican en memoria RAM sobre muestras en coma flotante normalizadas a [-1,1] y se ejecutan en el siguiente orden:

1. **Decodificación y validación.** Cuando el audio proviene de un contenedor con compresión con pérdida (por ejemplo, MP3, M4A u OGG), el primer paso consiste en convertirlo a una representación sin compresión denominada PCM (Pulse

Code Modulation). Este proceso implica reconstruir cada muestra de la forma de onda como valores numéricos crudos que representan su amplitud, eliminando cualquier tipo de compresión. En este trabajo se implementó una estructura de software predefinida que integra herramientas, bibliotecas y convenciones o tambien conocido como *workframe* llamado `FFmpeg`, el cual es llamado desde Python mediante la librería `subprocess` y una ruta configurable (`FFMPEG_PATH`). Esta configuración permite ejecutar el programa externo de manera controlada desde el código y obtener audio PCM estandarizado. La elección de `FFmpeg` responde a su amplia compatibilidad con la mayoría de códecs y formatos de archivo, su capacidad para manejar archivos complejos y su remuestreador de alta calidad, garantizando así un resultado robusto y consistente, independientemente del origen o tipo de fuente de audio.

2. **Eliminación de DC offset.** Cuando una señal de audio ingresa, suele presentar un leve desplazamiento respecto al nivel cero (*DC offset*), originado por tolerancias del hardware o sesgos del sistema de captura. Para corregirlo, se calcula la media de todas las muestras y se resta dicho valor a cada muestra de la señal. Este procedimiento re-centra la forma de onda alrededor de cero, recuperando el margen dinámico (*headroom*), reduciendo clics o ruidos en procesos de edición y evitando saturaciones asimétricas (*clipping*) que afectan más a un semieje que al otro.
3. **Downmix a mono.** Si el sistema detecta un archivo con canales estéreo (izquierdo y derecho), ambos se combinan con igual ponderación para obtener un único canal monofónico. Mediante el workframe `FFmpeg`, este proceso se realiza estableciendo el parámetro `-ac 1`, que aplica una mezcla balanceada y ajusta automáticamente el nivel de salida para evitar saturación. La conversión a un solo canal "Mono" resulta conveniente porque elimina posibles desfases entre canales y simplifica el procesamiento posterior del reconocedor automático de voz (ASR), el cual requiere una fuente acústica única, limpia y centrada para un rendimiento óptimo.
4. **Normalización de nivel.** Para que todos los audios ingresen con un volumen comparable, se mide el nivel RMS (valor eficaz) del fragmento y se aplica una ganancia global para llevarlo a un objetivo predefinido, por ejemplo, $-20\,\text{dBFS}$ (decibelios relativos a la escala de punto fijo completa). Si al aumentar el nivel los picos superan el margen permitido, el cual habitualmente se fija entre -1 y $-3\,\text{dBFS}$, se aplica una *limitación suave* (*soft clipping/limiting*) para controlar los picos sin generar distorsión abrupta. Asimismo, se mantiene un margen de 1–3 *dB* de "true peak", que representa el pico real estimado entre muestras tras la reconstrucción de la señal, una medida más estricta que el pico por muestra y fundamental para prevenir saturaciones en conversiones o codificaciones posteriores.
5. **Remuestreo a 16,kHz.** Cuando la frecuencia de muestreo original difiere de

16 kHz, la señal se remuestrea para adecuarla al formato estándar. Este proceso consiste en interpolar nuevas muestras y filtrar las componentes de frecuencia que no pueden representarse después del cambio. Mediante el uso del workframe `FFmpeg`, se especifica el parámetro `-ar 16000`, cuyo remuestreador incorpora un *filtro anti–alias* que atenúa las frecuencias superiores a 8 kHz, evitando así el repliegue espectral (*aliasing*). Esta herramienta se emplea por su alta precisión, ya que fijar la frecuencia a 16 kHz constituye una práctica ampliamente adoptada en sistemas ASR de voz conversacional, al reducir el costo computacional sin comprometer la inteligibilidad de la señal.

6. **Cuantización y exportación.** En el caso de este trabajo para almacenar el resultado de forma compatible y eficiente, la señal se convierte a enteros de 16 bits. Este paso se llama cuantización. Cuando procede, se añade un ruido muy bajo y aleatorio (dither) que disfraza los errores de cuantización y evita que aparezcan artefactos. En FFmpeg se fija el códec -acodec pcm_s16le, que hace que el flujo quede exactamente en PCM lineal de 16 bits, muy compatible con bibliotecas y modelos como Whisper. Por último, se serializa en un contenedor WAV y se guardan metadatos básicos (ruta, duración, frecuencia de muestreo y número de canales), garantizando un archivo sin compresión y listo para las siguientes etapas del sistema.

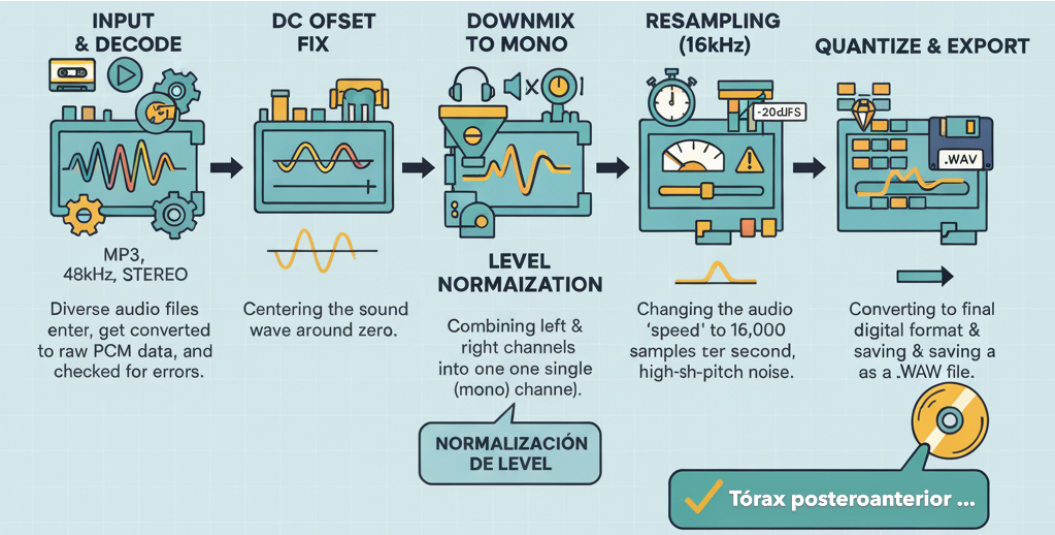
Caso de Uso

Dictado: *"tórax posteroanterior y lateral muestra un parénquima pulmonar"*.

1. Se recibe un archivo estéreo a 48 kHz en formato MP3. El módulo lo decodifica a PCM y verifica que los metadatos sean coherentes.
2. Se detecta un pequeño desplazamiento respecto a cero; se corrige eliminando el *DC offset* (esta desplazdo con respecto al origen).
3. Se combinan los canales izquierdo y derecho para obtener un único canal mono con la misma percepción de volumen.
4. El nivel RMS medido está por debajo del valor objetivo; por ello, se aplica una ganancia global que eleva la intensidad de la señal hasta -20 dBFS (valor que indica cuán cerca está la señal del nivel máximo que el sistema puede representar sin distorsión). Debido a la limitación suave, este aumento no produce *clipping*, es decir, no genera distorsión por exceder el límite máximo de amplitud permitido.
5. Se remuestrea de 48 kHz a 16 kHz usando filtrado anti–alias, el cual elimina las frecuencias altas que podrían causar *aliasing* es decir, la aparición de señales falsas o distorsionadas al reducir la frecuencia de muestreo, suprimiendo así el contenido por encima de 8 kHz.
6. El audio previamente estandarizado se cuantiza a 16 bits y se guarda como `.wav`. El fragmento que contiene "tórax posteroanterior y lateral..." queda homogéneo en formato, tasa y nivel, asegurando condiciones acústicas controladas y repro-

ducibles para el resto del flujo.

Figura 11. *Ilustración del Preprocesamiento de audio para estandarizar las entradas de audio.*



3.1.3. Módulo 3: Transcripción Automática Audio -Texto

Este módulo convierte el audio normalizado en texto, para lo cual fue implementado un motor ASR basado en Whisper-small (arquitectura encoder–decoder tipo Transformer, multilingüe). El proceso consiste en segmentar, caracterizar y convertir a términos textuales el audio de ingreso; este procedimiento es descrito a continuación:

1. **Inicialización del modelo.** Se implementó el modelo `Whisper`, utilizando en este trabajo la variante `small`, la cual requiere menores recursos computacionales. Este modelo permite disponer de un flujo ASR en español para generar una transcripción inicial (*greedy*) y activar posteriormente la opción de *beam search*. Se eligió este modelo debido a su robustez frente a entornos ruidosos y su adecuado equilibrio entre calidad de transcripción y latencia.
2. **Extracción de características.** Para la extracción de las características se emplea exclusivamente el audio normalizado con anterioridad. Este preprocesamiento garantiza que el modelo `small` de Whisper reciba una señal uniforme y óptima para la extracción de su representación log–Mel espectro–temporal interna, sin necesidad de cálculos adicionales manuales de espectrogramas.
3. **Segmentación temporal.** Se define un tamaño de ventana fijo, corto y adap-

table al proceso de tokenización, de modo que cada ventana genera uno o más *segmentos* con marcas de tiempo de inicio y fin. Se aplican solapes entre ventanas para minimizar pérdidas de información en los bordes. Whisper procesa cada ventana y devuelve segmentos con sus correspondientes palabras, marcas temporales y puntajes de confianza. Estos puntajes permiten identificar tokens de baja confianza que pueden ser posteriormente corregidos por el sistema de corrección clínica y médica, asegurando así una transcripción más precisa y confiable.

4. **Decodificación.** Para cada segmento, el decodificador genera la hipótesis de texto mediante un proceso de búsqueda autoregresiva. Se implementaron los métodos *greedy* y *beam search*, permitiendo al decodificador explorar múltiples posibilidades y seleccionar la más probable. Esta estrategia se utilizó para disminuir errores fonéticos, especialmente en terminología clínica, mejorando así la precisión de la transcripción.

5. **Alineación a nivel de palabra.** Los segmentos de audio previamente definidos mediante marcas temporales (tokens) se agrupan en palabras siguiendo las reglas del tokenizador interno de *Whisper*. Para cada palabra se registra: el texto, el instante de inicio, el instante de fin y un puntaje de confianza. En este trabajo, se aplicó un enmascaramiento a las palabras con confianza inferior a 0.65 utilizando la etiqueta `[MASK]`, lo que permitió posteriormente procesar dichos espacios mediante un modelo de lenguaje enmascarado para su corrección.

6. **Corrección léxica con RoBERTa (Masked-LM).** Para lograr la corrección léxica se uso la librería `transformers` en conjunto con el model `roberta-base-biomedical-clinical-es`, una variante optimizada de BERT de tipo *encoder-only* entrenada con aprendizaje de lenguaje enmascarado (*Masked-LM*) y mascara dinámica *dynamic masking*, capaz de adaptarse a las necesidades contextuales del texto. Esta versión está especializada en español biomédico, reforzando la cobertura de terminología clínica y abreviaturas. Se implementaron las librerías `transformers` y `PyTorch`, necesarias para el funcionamiento del modelo. Las palabras con confianza baja ($< \tau = 0{,}65$) se sustituyen por la etiqueta `[MASK]` y se aplica el flujo `fill-mask` con un parámetro `top_k` configurable, generando candidatos ordenados por probabilidad. La selección final pondera la probabilidad del modelo con la similitud ortográfica y fonética; si la ventaja sobre el segundo candidato no supera un umbral, se conserva la palabra original para evitar sustituciones inapropiadas. Este enfoque permite corregir con alta precisión errores típicos del ASR, como homófonos, tildes o confusiones de siglas, sin introducir contenido nuevo. Además, reduce la propensión a alucinaciones frente a modelos generativos, presenta baja latencia y costo computacional, y opera completamente en local, garantizando la privacidad de los datos clínicos.

7. **Composición de la salida.** Se escriben las hipótesis generadas por el modelo: T1 obtenida mediante *greedy* y T2 producida con *beam search* seguida del proceso de corrección, integrando así la transcripción final más precisa.

Caso de Uso

Dictado: *"tórax posteroanterior y lateral muestra un parénquima pulmonar"*.

1. El audio se convierte a espectrograma log–Mel y se procesa en una ventana única (duración corta), de la cual el modelo extrae un segmento.
2. La decodificación *greedy* produce una primera hipótesis coherente. La búsqueda con haz (*beam search*) explora alternativas y consolida la misma frase, corrigiendo detalles ortográficos (acentos) según la probabilidad agregada de los tokens. En otras palabras, la decodificación *greedy* elige palabra por palabra la opción más probable en cada paso, mientras que la búsqueda con haz considera varias posibles frases al mismo tiempo y selecciona la que tiene mayor coherencia global según el modelo.
3. El módulo devuelve el texto del segmento y, para cada palabra, su alineación temporal y confianza estimada. Un resultado ilustrativo es:

Palabra	**Inicio (s)**	**Fin (s)**	**Confianza**
tórax	0.10	0.55	0.96
posteroanterior	0.56	1.35	0.92
y	1.36	1.45	0.99
lateral	1.46	2.00	0.95
muestra	2.01	2.35	0.94
un	2.36	2.45	0.99
parénquima	2.46	2.85	0.88
pulmonar	2.86	3.10	0.93

4. Se debe notar que *parénquima* puede presentar una confianza ligeramente menor al tratarse de un término especializado; sin embargo, queda correctamente transcrito gracias a la evidencia acústica y al idioma fijado (en este caso español). El módulo entrega así una transcripción continua con marcas temporales y puntajes por palabra, lista para análisis posterior.

3.1.4. Módulo 4: Revisión semántica en el contexto de Radiología

Este módulo se encarga de refinar la transcripción generada por Whisper, transformándola en un texto clínicamente coherente y con terminología estandarizada. El proceso se organiza en dos sub-etapas complementarias: (i) corrección léxica mediante RoBERTa biomédico enmascarado y (ii) revisión semántica con el modelo Gemma 2 IT ajustado al español clínico. La entrada consiste en una transcripción con marcas temporales y pun-

tajes de confianza por palabra; la salida es un texto corregido que incluye abreviaturas normalizadas y redacción acorde al registro radiológico.

1. **Detección de incertidumbre léxica.** Se analizan los puntajes de confianza por palabra y se identifican como *candidatas a corrección* aquellas con puntaje inferior a un umbral $\tau = 0{,}65$. Palabras consecutivas de baja confianza se agrupan en *spans* que son las palabras consecutivas de baja confianza y se amplían con una ventana contextual corta para no dividir términos compuestos.
2. **Enmascaramiento y generación de candidatos.** Cada *span* se reemplaza por un token de máscara y se consulta el modelo RoBERTa biomédico (*fill-mask*) para generar un conjunto de k candidatos contextualmente relevantes. Se aplican filtros que incluyen normalización de tildes, restricción a vocabulario clínico y validación de unidades y siglas.
3. **Tokenización interna.** Aunque la corrección se realiza a nivel de palabra, internamente los modelos operan con tokens de subpalabras. Esto implica que:
   - Una palabra puede descomponerse en 1–N tokens, especialmente términos largos o con acentos (*pósteroanterior*, *parénquima*).
   - Al marcar una palabra, se expande la selección para abarcar todos sus tokens, evitando cortes dentro de subpalabras.
   - Los candidatos generados por RoBERTa se reensamblan en palabras o frases y se ordenan según: (i) probabilidad agregada, (ii) similitud ortográfica/fonética y (iii) restricciones clínicas.
   - Se aplica normalización Unicode (Consortium, 2025) que consiste en: colapso de espacios múltiples, inserción de espacios tras comas y dos puntos, eliminación de espacios antes de signos de puntuación, capitalización de inicio de frase y cierre con punto final si faltara. Esto garantiza consistencia en caracteres acentuados y símbolos.
4. **Selección y parcheo.** Se asigna a cada candidato un puntaje compuesto que considera tres factores: la confianza que RoBERTa tiene en la palabra sugerida, la similitud ortográfica y fonética con la palabra original, y la corrección morfológica, es decir, la concordancia de género, número y uso correcto de mayúsculas. El sistema reemplaza la palabra original solo si la diferencia de puntaje entre el candidato principal y el siguiente es suficientemente significativa, minimizando así sustituciones incorrectas o innecesarias. La integración de la corrección se realiza utilizando los índices de palabra, de manera que se preservan todas las marcas de tiempo generadas por Whisper. Adicionalmente, cada cambio se documenta en un registro estructurado (JSON) que incluye la palabra original, el reemplazo aplicado, la razón de la corrección y el rango de tokens afectado, garantizando trazabilidad completa y auditoría del proceso.
5. **Revisión semántica instruida.** Se emplea un LLM tipo *instruct* local, car-

gado con `Gemma-2-IT` cuantizado en GGUF mediante `llama_cpp`. Se construyen *prompts* que combinan salidas greedy y beam search, con instrucciones de estilo y seguridad: no agregar hallazgos, homogeneizar abreviaturas y mantener español clínico. Se controla la decodificación mediante temperatura, top-p/top-k, penalización por repetición y *stop words*. Se elimina encabezado fijo [Transcripción Final :] para que la salida quede lista para su presentación sin posprocesado adicional.

6. **Empaquetado de la salida.** Se aplica un diccionario externo `Diccionario.json` que corrige términos con errores registrados con anterioridad por el operador, tanto *in-prompt* para el LLM como *post-hoc* sobre el texto final. Se ejecutan normalizaciones Unicode, es decir, se aplica colapso de espacios, ajuste de puntuación y capitalización, preservando marcas de tiempo y segmentación. Cada modificación se documenta en una lista serializable para auditoría, asegurando coherencia terminológica y reproducibilidad.

Caso de Uso

Dictado: *"tórax posteroanterior y lateral muestra un parénquima pulmonar"*.

1. **Detección.** Se marcan como candidatas "posteroanterior" (sin tilde) y, si apareciera, "parenquima" (sin tilde), por presentar confianza moderada.
2. **Enmascaramiento.** El texto local se transforma a: "tórax `[MASK]` y lateral muestra un `[MASK]` pulmonar", preservando el resto de palabras y sus tiempos.
3. **Candidatos para la primera máscara.** RoBERTa propone "pósteroanterior", "postero anterior", "postero–anterior". Se selecciona **"pósteroanterior"** por adecuación terminológica y ortográfica.
4. **Candidatos para la segunda máscara.** RoBERTa sugiere "parénquima", "parenquimatoso", "parénquimas". Se elige **"parénquima"** por concordancia morfosintáctica con "un ... pulmonar".
5. **Parcheo.** El texto queda: "tórax pósteroanterior y lateral muestra un parénquima pulmonar".
6. **Revisión semántica.** Gemma 2 IT fine tuneado ajusta puntuación y estilo sin alterar el contenido. Una salida típica es:

   > *"Radiografía de tórax en proyecciones pósteroanterior y lateral: se observa parénquima pulmonar."*

7. *Resultado y trazabilidad.* Se devuelve el texto corregido y el registro de cambios: {"posteroanterior" → "pósteroanterior" (tilde), "parenquima" → "parénquima" (tilde)}.

3.1.5. Módulo 5: Generación del informe final

En ésta etapa se transforma el texto previamente corregido en una transcripción de un informe clínico listo para uso, operando sobre tres ejes fundamentales: (i) *Corrección y ajuste ligero* para garantizar legibilidad y uniformidad, (ii) *serialización y entrega* segura del contenido, y (iii) *telemetría de interfaz* para reportar el estado del sistema en tiempo real. Todas estas operaciones se realizan en el hilo de interfaz con tareas no bloqueantes, preservando la *responsividad* de la GUI.

1. **Corrección y ajuste ligero** para este propósito se aplican reglas de normalización Unicode para unificar los caracteres con acentos (por ejemplo, $a + \acute{} \rightarrow \acute{a}$), colapso de múltiples espacios y tabulaciones a un solo espacio, ajuste del espaciado después de comas, dos puntos y punto y coma, eliminación de espacios indebidos antes de signos de cierre, capitalización del inicio de oración y adición automática de punto final cuando falta. Estas reglas se aplican sobre el texto final sin alterar el contenido clínico, evitando duplicidades, errores en comparaciones y artefactos de tokenización, mientras se asegura la legibilidad y consistencia tipográfica.
2. **Presentación en la GUI.** El texto se proyecta en un editor de sólo lectura con soporte para selección y conteo. Se utiliza un *buffer* de salida con *append incremental* y *throttling* por lotes, permitiendo actualizaciones fluidas sin bloquear el hilo principal. Se preserva el foco y se implementa *scroll anchoring* para evitar que el cursor y la selección del usuario se desplacen de manera inesperada durante la inserción de texto.
3. **Copia al portapapeles.** El texto puede exportarse directamente al portapapeles del sistema en formato `UTF-8`, conservando diacríticos y garantizando compatibilidad con HIS, PACS o editores externos. Se normalizan los finales de línea (`LF`↔`CRLF`) según la plataforma, asegurando que el contenido pegado respete el entorno destino.
4. **Guardado en `.txt`.** La persistencia en disco se realiza en texto plano `UTF-8` mediante escritura atómica: se vuelca primero a un archivo temporal en el mismo directorio, seguido de `flush`/`fsync` y `rename`/`replace`, garantizando que el archivo final no quede corrupto ante fallos. Se aplica nomenclatura con sello temporal ISO 8601 (`YYYY-MM-DD_hh-mm-ss`) para asegurar unicidad y facilitar auditoría y archivo clínico.
5. **Limpieza controlada.** Se restablece el estado del editor sin afectar archivos persistidos. Esto incluye la limpieza del *buffer* de salida y el reinicio de contadores de vista, permitiendo iniciar nuevas transcripciones sin arrastre de contenido previo. Se solicita confirmación cuando hay cambios sin guardar, evitando pérdidas accidentales de información crítica.
6. **Indicadores de estado y telemetría.** Se visualizan periódicamente métricas

de CPU, RAM y VRAM junto con el progreso del proceso. Se realiza muestreo con temporizador a una frecuencia limitada ($\leq 4\,\text{Hz}$) para no saturar el hilo de interfaz. La agregación por ventana (*debounce*) y la detección de picos permiten mostrar tendencias estables y alertas relevantes, evitando confusión por señales ruidosas.

7. **Trazabilidad y registro.** mantiene una bitácora cronológica de eventos y acciones del usuario, implementando niveles de registro (`INFO/WARNING/ERROR`) y un doble *handler* (archivo y consola). Cada acción (copiar, guardar, limpiar) y transición de estado se registra con sello temporal, ruta de archivo y resumen, facilitando auditoría clínica y depuración. Los mensajes de error y las sesiones se etiquetan de forma que los incidentes sean reproducibles y el soporte técnico pueda actuar de manera eficiente.

Caso de Uso

Dictado de referencia: *"tórax posteroanterior y lateral muestra un parénquima pulmonar"*.

1. El módulo recibe el texto ya corregido ortográfica y terminológicamente. Aplica normalización tipográfica: corrige espaciado, añade el punto final si falta y asegura capitalización inicial.
2. El panel de informe muestra:

   > *Radiografía de tórax en proyecciones pósteroanterior y lateral: se observa parénquima pulmonar.*

   con conteo de *14 palabras / 102 caracteres* (valores ilustrativos).
3. El usuario pulsa *Copiar* y el informe queda disponible en el portapapeles (UTF-8), listo para pegarse en el sistema de informes hospitalario.
4. El usuario pulsa *Guardar*, elige `informe_torax_2025-10-05_143210.txt` y el sistema realiza una escritura atómica. La GUI notificará de esta acción.
5. Finalmente, con *Limpiar* se deja el panel vacío para el siguiente caso. Durante todo el proceso, las tarjetas de CPU/RAM/VRAM actualizan su lectura sin afectar la interacción.

3.2. Interfaz gráfica de usuario (GUI)

El sistema propuesto se ejecuta en una única interfaz gráfica, como se ilustra en la Figura 12, y está organizada mediante un *layout* de paneles fijos para reducir la carga cognitiva del usuario. La GUI se implementó con `PyQt6`, utilizando contenedores como `QVBoxLayout`, `QHBoxLayout`, `QSplitter` y `QGroupBox` para estructurar secciones, logrando una disposición clara y escalable. La parte superior exhibe *telemetría del sistema*, el panel central se divide en *Paso 1: Selección/Grabación de Audio*, *Paso 2:*

Configuración y *Panel de resultados*, mientras que la franja inferior agrupa las *acciones del informe*.

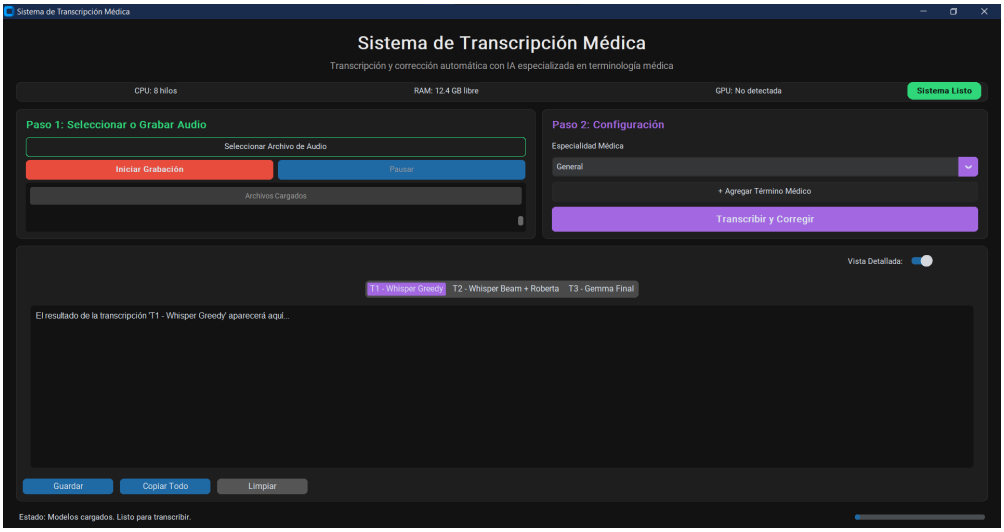
La GUI opera mediante una máquina de estados (*inactivo, grabando, pausado, procesando, listo, error*) que habilita o deshabilita controles según la transición, evitando combinaciones inválidas (p. ej., procesar sin audio). Las tareas prolongadas (grabación, conversión de audio, transcripción y corrección) se ejecutan en hilos `QThread`, lo que permite que el `mainloop` solo gestione la renderización y recepción de eventos. La comunicación de progreso y resultados entre hilos y la interfaz se realiza mediante señales (`pyqtSignal`) y `QTimer`.

Para la interacción con el usuario se integraron:

- `QFileDialog` para apertura y guardado de archivos.
- `QMessageBox` para notificaciones informativas.
- `QMenu` y menús contextuales para acciones rápidas, incluyendo la adición de términos médicos seleccionados al diccionario.

Los botones de control de audio están conectados con el módulo `AudioRecorder`, que utiliza `sounddevice` y `NumPy` para capturar y almacenar la señal en formato PCM 16-bit/16 kHz mono.

Figura 12. *Interface gráfica principal del sistema propuesto.*



3.2.1. Componentes y Funcionalidades clave

- **Barra de estado superior:** Implementada con `QFrame` y `QLabel`, presenta núcleos de CPU, memoria libre y disponibilidad de GPU. El indicador de *estado del sistema* refleja la máquina de estados. La actualización periódica mediante `QTimer` (~ 2 s) evita bloqueos de la interfaz.

- **Paso 1: Selección/Grabación de Audio:** El usuario puede cargar un archivo con `QFileDialog` o iniciar grabación mediante `AudioRecorder`. Los botones *Iniciar*, *Pausar* y *Detener* se habilitan según el estado actual. La señalización de progreso y duración se refleja en un `QLabel`.
- **Paso 2: Configuración:** Se ofrece un `QComboBox` para elegir la especialidad médica y un botón para agregar términos al diccionario, que abre un diálogo modal (`QDialog`). La acción *Transcribir y Corregir* inicia un hilo de transcripción y corrección, bloqueando entradas incompatibles durante el procesamiento.
- **Panel de resultados:** Un `QTextEdit` personalizado muestra la transcripción en tiempo real. Incluye menú contextual para cortar, copiar, pegar, seleccionar todo y agregar términos al diccionario. La conmutación entre vistas intermedias y finales se realiza mediante botones.
- **Acciones del informe:** Botones para *Guardar*, *Copiar* y *Limpiar* la transcripción. La operación de guardado emplea escritura atómica y soporte UTF-8; el portapapeles nativo asegura compatibilidad con HIS/PACS. Todas las acciones notifican al usuario y registran eventos para trazabilidad.

3.2.2. Caso de Uso

Dictado de referencia: *"tórax posteroanterior y lateral muestra un parénquima pulmonar"*.

1. **Preparación.** La GUI arranca en estado *inactivo* con el indicador *Sistema listo*. El usuario selecciona *Especialidad: General* y añade, si lo desea, un término (*p. ej.*, "pósteroanterior") al diccionario mediante el cuadro *+ Agregar término*. La barra superior muestra: *CPU: 8 hilos, RAM libre: 12.4 GB, GPU: No detectada*.
2. **Entrada de audio.** El usuario elige un archivo o inicia la grabación. Al haber un ítem activo, el botón "Transcribir y Corregir"se habilita y la lista *Archivos cargados* refleja la selección.
3. **Ejecución y visualización.** Al pulsar "Transcribir y Corregir", la interfaz cambia a estado *procesando*. Las etiquetas superiores del panel de resultados se activan:

- `T1 - Whisper Greedy`: aparece en primer lugar un borrador temprano del dictado.
- `T2 - Whisper Beam + Roberta`: se muestra una versión depurada, manteniendo resaltadas las palabras ajustadas.
- `T3 - Gemma Final`: se presenta el informe estilizado en registro radiológico.

4. **Entrega del informe.** Con el texto final visible, el usuario pulsa Çopiar todo"para llevarlo al portapapeles o "Guardar"para guardar el texto en archivo `.txt` con nombre sugerido (*fecha_hora*). La barra de estado confirma la operación y el sistema vuelve a *listo*.
5. **Ciclo siguiente.** El botón "Limpiar"vacía el panel de resultados y reinicia contadores visuales. El historial de acciones (copiado/guardado/limpieza) queda anotado para auditoría.

3.3. Entrenamiento del Modelo

En esta sección se detalla el ajuste fino del modelo conversacional *Gemma-2B-IT* para la corrección de informes radiológicos en español, empleando una estrategia eficiente en memoria mediante QLoRA. QLoRA combina dos conceptos clave: primero, los pesos del modelo base se almacenan en cuatro bits utilizando la cuantización NF4; segundo, se añaden adaptadores LoRA, módulos pequeños y entrenables que se insertan en capas específicas de la red, permitiendo actualizar únicamente estos parámetros mientras los pesos originales permanecen congelados. Esta técnica reduce significativamente el consumo de memoria, mantiene una calidad cercana a la precisión completa y permitió entrenar un modelo de aproximadamente dos mil millones de parámetros usando una sola GPU A100 de 40 GB sin comprometer la estabilidad ni incurrir en sobreajuste.

3.3.1. Arquitectura base y ajuste fino

Gemma-2B-IT es un modelo tipo Transformer orientado a generación de texto y a instrucciones conversacionales. Para adaptarlo a la terminología radiológica, se emplearon pares *entrada–salida* reales, donde se penalizan los errores palabra por palabra de manera autoregresiva durante el entrenamiento.

3.3.2. Justificación del uso de LoRA y QLoRA

La técnica LoRA (*Low-Rank Adaptation*) añade adaptadores de bajo número de parámetros en proyecciones de atención y en los bloques densos (MLP), permitiendo aprender las correcciones necesarias sin modificar los pesos originales. Esto reduce el número de parámetros a actualizar, ahorra memoria, acelera el entrenamiento y actúa

como regularización al restringir las modificaciones a un subespacio dirigido (Hu et al., 2021).

QLoRA extiende esta estrategia cuantizando los pesos base en cuatro bits mediante NF4, mientras solo los adaptadores LoRA se entrenan en precisión mixta. La cuantización NF4 preserva mejor la distribución estadística de los pesos frente a esquemas de cuatro bits, manteniendo la estabilidad del entrenamiento con un consumo de memoria reducido (Dettmers et al., 2023).

3.3.3. Flujo técnico del entrenamiento

Se utilizaron los siguientes componentes:

- **`transformers`**: carga de `Gemma-2B-IT`, tokenización, control de longitud, colación de lotes y cómputo de la pérdida autoregresiva. Esta librería permite la integración de adaptadores LoRA (`peft`) y la cuantización NF4 (`bitsandbytes`).
- **`peft`**: inyecta adaptadores LoRA en capas específicas del modelo, manteniendo congelados los pesos originales y gestionando su ciclo de vida y fusión opcional.
- **`bitsandbytes`**: habilita cuantización NF4 a cuatro bits y estados de optimizador en ocho bits con paginación a RAM, permitiendo entrenamiento eficiente en VRAM limitada.
- **`accelerate`**: coordina distribución de modelos y tensores, precisión mixta, escalado automático de pérdidas y acumulación de gradientes, simplificando la configuración reproducible.
- **Exportación a `GGUF`**: el modelo base junto con los adaptadores entrenados se fusiona y se convierte en formato `GGUF` para inferencia local eficiente con `llama_cpp`.

3.3.4. Conjunto de datos

Se empleó un esquema supervisado con pares *prompt–target*:

- `Entrenamiento_ampliado.json`: 90 % entrenamiento, 10 % validación.
- `Evaluación.json`: exclusivo para evaluación objetiva.

Cada ejemplo se encapsula en la plantilla conversacional de *Gemma-2B-IT*:

```
<start_of_turn>user Corrige este informe radiológico:
{prompt}
<end_of_turn>
<start_of_turn>model
{target}
<end_of_turn>
```

Se aplicó normalización Unicode NFC, limpieza de espacios, control de longitud por tokens y verificación de información sensible, manteniendo la representatividad por tipo de estudio en la partición entrenamiento/validación.

3.3.5. Estrategia LoRA–QLoRA y recursos

- Cuantización NF4 a cuatro bits para los pesos base.
- Adaptadores LoRA entrenables insertados en atención y MLP; el resto del modelo permanece congelado.
- Precisión mixta con activaciones y gradientes en 16 bits, estados del optimizador en ocho bits con paginación a RAM.
- Bibliotecas: `transformers`, `peft`, `bitsandbytes`, `accelerate`, `datasets`, `sacrebleu` y `rouge-score`.
- Hardware: GPU A100 de 40 GB; artefactos y métricas guardados externamente en formato JSONL.

3.3.6. Hiperparámetros

Tabla 3-1: Hiperparámetros del ajuste fino de Gemma-2B-IT

Parámetro	Valor
Lote efectivo	8 (mini-lote 2 con acumulación 4)
Épocas	3
Tasa de aprendizaje	2e-5
LoRA (r, alpha, dropout)	8, 32, 0.05
Cuantización	4 bits, NF4
Optimizador	`paged_adamw_8bit`
Planificador de LR	Coseno con 10 % calentamiento
Longitud máxima	2 048 tokens (empaquetado)
Semilla	42

3.3.7. Procedimiento resumido

1. Preparación de datos: normalización, limpieza y empaquetado en la plantilla conversacional.

2. Inicialización: carga del modelo base cuantizado y adición de adaptadores LoRA.
3. Entrenamiento: acumulación de gradientes, optimizador en ocho bits, precisión mixta y planificador coseno.
4. Selección de punto de control: según ROUGE-L en validación.
5. Fusión y exportación: adaptadores integrados en el modelo base y conversión a GGUF para inferencia local.

3.3.8. Evaluación del sistema

La evaluación tuvo como propósito cuantificar el desempeño del modelo *Gemma-2B-IT* tras el ajuste fino en la corrección de informes radiológicos en español. Se buscó medir con precisión la capacidad del modelo para generar textos corregidos que mantuvieran la terminología médica, la coherencia lingüística y la fidelidad respecto a la referencia esperada.

Diseño metodológico

El modelo se evaluó de manera supervisada usando el conjunto independiente denominado `Evaluación.json` con un total de 100 informes, que contiene pares *prompt–target* no vistos durante el entrenamiento. Cada *prompt* representa un informe con errores ortográficos, gramaticales o terminológicos simulados, mientras que el *target* es la versión correctamente redactada y validada.

Durante la implementación, se utilizó decodificación con temperatura baja (`temperature = 0.1`) y muestreo restringido (`top_p = 0.9`) para eliminar la variabilidad aleatoria y obtener resultados reproducibles. Las salidas generadas por el modelo se compararon directamente con los textos de referencia, empleando métricas de similitud textual ampliamente utilizadas en la evaluación de modelos de lenguaje: BLEU y ROUGE.

Métrica BLEU

BLEU (*Bilingual Evaluation Understudy*) mide el grado de coincidencia entre los n-gramas (secuencias de una o más palabras) del texto generado y los del texto de referencia. En este trabajo se calcularon los promedios de BLEU-1 a BLEU-4 para evaluar desde la precisión léxica individual hasta la coherencia de frases completas.

Una puntuación alta de BLEU indica que el modelo logra reproducir con fidelidad las palabras y combinaciones esperadas, reflejando su capacidad para corregir errores sin alterar el contenido semántico del informe original. Esta métrica es especialmente útil en tareas de corrección, ya que penaliza sustituciones y omisiones no justificadas, premiando la exactitud literal del texto corregido.

Métrica ROUGE

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) evalúa la similitud entre el texto generado y el de referencia desde una perspectiva de recuperación de información. A diferencia de BLEU, ROUGE valora más la cobertura del contenido relevante que la precisión exacta.

En esta evaluación se calcularon las variantes ROUGE-1, ROUGE-2 y ROUGE-L:

- **ROUGE-1:** mide el solapamiento de unigramas, es decir, la coincidencia palabra a palabra.
- **ROUGE-2:** evalúa la secuencia de bigramas, reflejando la fluidez local entre palabras.
- **ROUGE-L:** se basa en la subsecuencia común más larga (*Longest Common Subsequence*), lo que permite medir la preservación del orden y la coherencia global del texto.

De esta manera, mientras BLEU analiza la precisión de los términos generados, ROUGE mide la completitud y coherencia del texto final. El uso combinado de ambas métricas proporciona una visión integral del rendimiento del modelo: BLEU confirma la exactitud terminológica y ROUGE evalúa la continuidad estructural y contextual.

Procedimiento de evaluación

Cada texto generado fue comparado con su referencia mediante las librerías `sacrebleu` y `rouge-score` (Post, 2018; Research/opensource, 2020). Ambas métricas se calcularon de forma independiente para cada ejemplo y posteriormente se promediaron sobre todo el conjunto de evaluación, garantizando representatividad estadística. Además, se analizaron los resultados por subcategoría de estudio (tórax, abdomen, cráneo, extremidades) para verificar consistencia entre distintos estilos de redacción.

En conjunto, esta metodología permitió una valoración cuantitativa, objetiva y reproducible del desempeño del modelo tras el ajuste fino, centrada en la precisión, cobertura y coherencia de las correcciones producidas.

4 Resultados

Esta sección presenta los resultados obtenidos al especializar Gemma-2B-IT para la corrección de informes radiológicos en español. Se describen la configuración experimental, el protocolo de evaluación, las métricas empleadas y los hallazgos cuantitativos y cualitativos, junto con una discusión crítica de su impacto, limitaciones y reproducibilidad.

4.1. Configuración experimental

4.1.1. Entorno computacional

Entrenamiento (Google Colab).

- **GPU**: NVIDIA A100 (arquitectura *Ampere*), memoria: 40 GB HBM2e, ancho de banda: ∼1,555 GB/s.
- **Tensor Cores**: 3.ª generación.
- **Precisiones soportadas**: FP64, FP32, TF32, BF16, FP16, INT8.

Pruebas de inferencia (equipo local).

- **Portátil**: AORUS 15P KD;
- **GPU**: NVIDIA RTX 3060 (6 GB VRAM);
- **CPU**: Intel Core i7;
- **RAM**: 16 GB.

4.1.2. Conjunto de Datos

Se usaron un flujo de 20000 pares de informes (entrada → referencia) a partir de informes radiológicos en español. Cada entrada se generó introduciendo de manera sintética errores típicos (errores generados) de sistemas ASR genéricos o no adaptados al dominio radiológico como lo son: la omisión de palabras, sustituciones ortográficas (*pósteroanterior* → *posteroanterior*), confusión de siglas/unidades o puntuación, entre otros varios errores. Para cada entrada se conservó su *contraparte correcta* (referencia), garantizando una señal de entrenamiento alineada.

4.1.3. Estrategia de ajuste fino (QLoRA)

El modelo conversacional Gemma-2B-IT se especializó con QLoRA, que combina cuantización de 4 bits (NF4) en los pesos congelados con adaptadores LoRA entrenables de bajo rango:

- **LoRA (Low-Rank Adaptation).** En lugar de volver a entrenar todos los pesos del modelo, LoRA añade un "adaptador" pequeño en capas clave (atención y MLP) mientras deja el modelo base *congelado*. Ese adaptador se describe como una corrección ΔW de *bajo rango* (dos matrices pequeñas que, al multiplicarse, forman la actualización). En la práctica: (i) se entrenan muy pocos parámetros, (ii) se usa menos VRAM y (iii) se reduce el riesgo de sobreajuste porque el cambio permitido al modelo está "acotado" a un subespacio pequeño. En inferencia, el adaptador se suma al peso original ($W' = W + \Delta W$) sin añadir latencia apreciable (Hu et al., 2021).
- **QLoRA.** Combina LoRA con *cuantización* para ahorrar memoria. Los pesos del modelo base se mantienen fijos y se almacenan en 4 bits usando el formato NF4, que comprime bien sin perder información esencial. Mientras tanto, *sólo* se entrenan los adaptadores LoRA en 16 bits (BF16/FP16) para conservar estabilidad numérica. Así, QLoRA permite afinar modelos grandes en una sola GPU: los pesos cuantizados ocupan poco, y el cómputo pesado recae únicamente en el pequeño adaptador entrenable (Dettmers et al., 2023).

4.1.4. Hiperparámetros aplicados

(coherentes con el diseño previo del sistema):

- **épocas**= 3
- **LR** = 2×10^{-5}
- ***batch* efectivo**= 8
- $(r, \alpha, p_{\mathbf{drop}}) = (8,\ 32,\ 0{,}05)$

Esta combinación equilibró *uso de memoria* y *generalización*, evitando oscilaciones y sobreajuste.

4.1.5. Decodificación en inferencia

Para *todas* las comparaciones se mantuvo fija la misma configuración de generación del modelo: temperatura 0,7 (controla el grado de aleatoriedad en la selección de la siguiente palabra; valores más altos generan respuestas más diversas y valores más

bajos, más deterministas), *top-k* = 50 (la siguiente palabra se elige sólo entre las 50 más probables) y *top-p* = 0,95 (muestreo por núcleo que limita la elección al conjunto mínimo de palabras cuya probabilidad acumulada alcanza el 95 %). Además, se aplicó una *detokenización* homogénea (reconstrucción del texto continuo a partir de *tokens* internos, eliminando marcadores y reinsertando espacios) y una normalización Unicode en forma canónica NFC el cual unifica caracteres equivalentes; por ejemplo, la "á" precompuesta frente a "a" + acento antes del cómputo de métricas, con el fin de evitar sesgos debidos a variantes de codificación o espaciado.

4.1.6. Protocolo de evaluación

Se utilizó un **conjunto de prueba** de 100 informes, independiente del material de entrenamiento. Cada ejemplo consiste en un par *prompt–target*: el *prompt* emula el dictado con errores (ortografía, tipografía, puntuación y terminología), y el *target* es su versión corregida. Se evaluaron dos sistemas bajo la misma configuración de inferencia:

1. **Modelo base**: Gemma-2B-IT sin especialización específica para radiología.
2. **Modelo ajustado**: Gemma-2B-ITFT especializado con QLoRA y la estructura descrito.

Para asegurar imparcialidad, se mantuvieron constantes el preprocesamiento, el *prompting* y los límites de longitud de entrada/salida. Las salidas se normalizaron a minúsculas con preservación de diacríticos para el cómputo de *n-gramas*.

El desempeño se cuantificó con BLEU y ROUGE (`rouge_score`), métricas *estándar* para tareas de corrección/reformulación textual.

- **BLEU:** Mide el solapamiento de n-gramas entre hipótesis y referencia con penalización por brevedad. Es sensible a errores ortográficos, de tokenización y diacríticos, por lo que captura bien la precisión tipográfica.
- **ROUGE-1 y ROUGE-2:** Evalúan cobertura de unigramas y bigramas, respectivamente. ROUGE-1 refleja precisión léxica y ROUGE-2 la fluidez local al considerar parejas contiguas de palabras clínicas (*p. ej.*, *patrón reticulonodular*).
- **ROUGE-L:** Basado en la *Longest Common Subsequence*, aproxima la coherencia global y el respeto de la estructura del informe (hallazgo → localización → característica).

4.1.7. Resultados obtenidos

La Tabla **4-2** resume los incrementos absolutos y relativos del modelo ajustado respecto del base.

Los aumentos son tal como se puede apreciar en la Figura 13 teniendo mejorías en todas las métricas y las cuales se interpretan en mejoras cualitativas concretas:

Tabla 4-1: Comparación de métricas entre el modelo base y el modelo ajustado (n = 100).

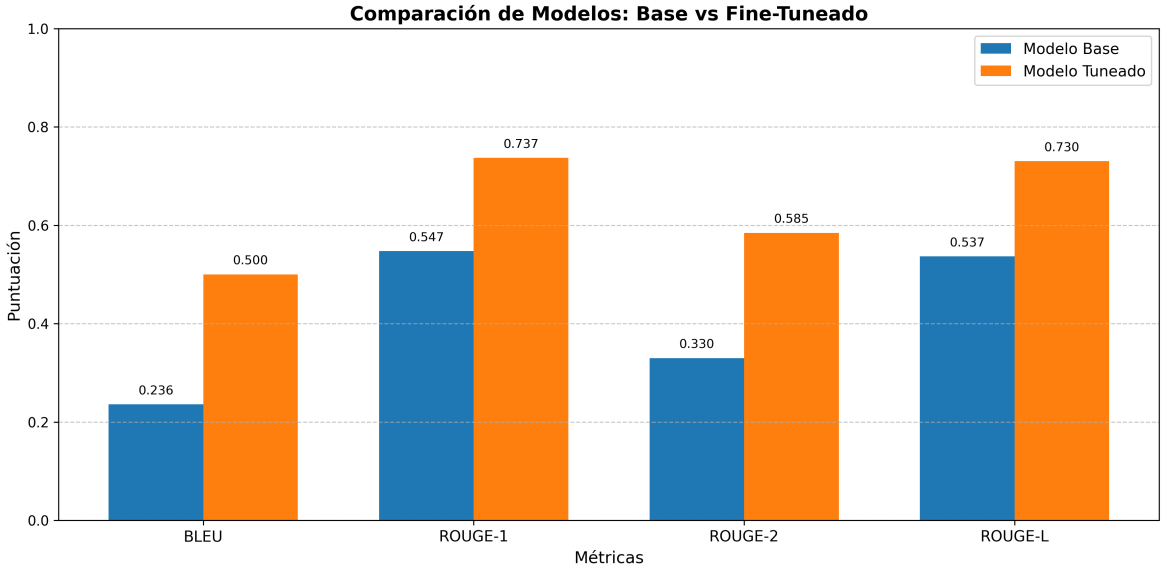
Métrica	**Modelo base**	**Modelo ajustado**	**Mejora (%)**
BLEU (0–100)	23.60	**50.02**	+112.0
ROUGE-1	0.547	**0.737**	+34.6
ROUGE-2	0.330	**0.585**	+77.3
ROUGE-L	0.537	**0.730**	+36.0

Tabla 4-2: Incrementos absolutos y relativos del modelo ajustado respecto al base.

Métrica	**Incremento absoluto**	**Incremento relativo**
BLEU (puntos)	+26.42	+112.0 %
ROUGE-1	+0.190	+34.6 %
ROUGE-2	+0.255	+77.3 %
ROUGE-L	+0.193	+36.0 %

- **Precisión léxica y ortográfica** (BLEU, ROUGE-1). Disminuye la tasa de tildes omitidas y grafías erróneas (*parénquima*, *pósteroanterior*), y se regulariza la capitalización. La ganancia en BLEU evidencia que el modelo corrige con mayor fidelidad *n-gramas* exactos sensibles a acentuación.
- **Fluidez local** (ROUGE-2). Se recuperan bigramas clínicos frecuentes (*engrosamiento broncovascular*, *patrón reticulonodular*), lo que mejora la cohesión inmediata entre términos y reduce inserciones/omisiones de conectores.
- **Estructura global** (ROUGE-L). Se preserva mejor la secuencia típica del informe (hallazgo → localización → características), con puntuación más estable y cierres de enunciado consistentes.

Figura 13. *Comparación gráfica de las métricas promedio entre el modelo base y el modelo ajustado.*



4.1.8. Relación con el diseño (QLoRA + LoRA).

El uso de LoRA (rango $r = 8$, $\alpha = 32$, *dropout* $= 0{,}05$) concentró la capacidad de adaptación en las proyecciones relevantes del Transformer, limitando el número de parámetros entrenables y mitigando el sobreajuste. La cuantización NF4 de QLoRA redujo drásticamente memoria sin sacrificar estabilidad al entrenar los adaptadores en 16 bits; ello permitió explorar $\geq$3 épocas con *batch* efectivo adecuado, mejorando convergencia sin oscilaciones.

5 Discusión

Los resultados cuantitativos evidencian que el ajuste fino de Gemma-2B-IT con QLoRA y adaptadores LoRA produce mejoras sustanciales sobre el modelo base (Tabla **4-1** y Fig. 13). En BLEU el incremento absoluto es de 26.4 puntos (112 % relativo), y en ROUGE las ganancias van de 34.6 % (ROUGE-1) a 77.3 % (ROUGE-2), con un 36.0 % en ROUGE-L. A continuación se discuten, desde una perspectiva técnica, los factores que explican estas mejoras, su alineación con la literatura, las limitaciones observadas y las implicaciones prácticas para entornos clínicos.

5.0.1. Factores que explican las ganancias

1. **Adaptación de dominio orientada por instrucciones.** El formato de *prompt–target* y la forma en trabaja el modelo hacen que el sistema use el tipo de lenguaje y las frases que son comunes en los informes de radiología (orden “hallazgo → localización → características”). Este enfoque restringe la variedad de respuestas posibles del sistema (reduciendo la "dispersión del decodificador"), lo cual se traduce en una mejora medible de la coherencia general (indicada por la métrica ROUGE-L) y de la pertinencia y riqueza del vocabulario utilizado (indicada por ROUGE-1) (Wei et al., 2024).
2. **Especialización eficiente con LoRA–QLoRA.** El enfoque de Especialización eficiente con LoRA y QLoRA permite adaptar un modelo de lenguaje grande a un dominio específico, como la radiología, sin perder su conocimiento original. Esto se logra congelando la mayor parte del modelo (los "pesos base"), que está reducida a solo 4 bits para ahorrar memoria, preservando así el conocimiento general del mundo. Al mismo tiempo, se introducen y entrenan pequeñas estructuras llamadas matrices de bajo rango (con un rango $r = 8$) únicamente en las partes de ”atención” del modelo. Estas matrices son las que aprenden los patrones de lenguaje y el vocabulario específicos del dominio, como las frases clínicas exactas (”patrón reticulonodular”). Este diseño tiene dos beneficios clave: evita que el modelo ”olvide” su información original y asegura una capacitación estable usando menos recursos computacionales. La mejora significativa en la métrica ROUGE-2 confirma que el modelo ha aprendido a usar estas combinaciones de dos palabras (bigramas) relevantes para el contexto clínico, lo cual es una señal de que esta especialización ha sido efectiva al ajustar selectivamente cómo el modelo ”presta atención” a la información.

3. **Normalización ortográfica y tokenización robusta.** El español clínico implica tildes, guiones y compuestos (p. ej., *pósteroanterior*). Al transformar cadenas de texto a una única representación canónica o tambien llamada normalización Unicode y a mantener la consistencia en la plantilla conversacional utilizada por el modelo, se logra reducir significativamente los errores en la segmentación de las palabras en subpalabras (un proceso interno del modelo de lenguaje). Esta precisión en el procesamiento del texto tiene un impacto directo y positivo en la calidad de la traducción o generación textual medida por métricas como BLEU (que es muy sensible a cualquier discrepancia ortográfica) y ROUGE-1.
4. **Regularización por diseño.** El método de entrenamiento empleado incorpora una serie de técnicas como lo son la tasa de aprendizaje (LR) moderada, un periodo inicial de incremento suave de esta tasa (warm-up del 10 %), una disminución gradual siguiendo una curva de coseno (cosine decay), y la aplicación de una técnica de abandono o regularización específica para el ajuste fino (dropout LoRA=0.05), con el fin de actuar como un mecanismo de regularización por diseño". Este conjunto de estrategias está concebido para mitigar los efectos desestabilizadores que pueden surgir al comprimir el modelo a solo 4 bits (cuantización), promoviendo una convergencia estable y fluida. Al hacerlo, el proceso optimiza el balance entre la precisión general del contenido o las palabras clave (evaluada por métricas como BLEU y ROUGE-1) y la calidad en la secuencia y conexión de las frases a nivel local (medida por ROUGE-2), asegurando que el modelo resultante sea robusto y produzca resultados coherentes.
5. **Decodificación conservadora.** La estrategia de "Decodificación conservadora" consiste en generar texto de manera controlada para asegurar que las respuestas del modelo de inteligencia artificial sean precisas y fieles a la información original (el target), en lugar de ser creativas o variadas. Para lograr esto, se ajustan parámetros clave durante el proceso de inferencia (cuando el modelo produce la respuesta): una Temperatura de 0.7 introduce una pequeña dosis de aleatoriedad para evitar repeticiones, mientras que los ajustes top-k = 50 y top-p = 0,95 limitan la selección de palabras a un conjunto muy probable y de alta confianza, restringiendo así las opciones del modelo para que solo elija términos que tienen una alta probabilidad de ser correctos. En el ámbito de los textos técnicos, este enfoque es crucial porque reduce la incertidumbre o la entropía en la selección de palabras, forzando al modelo a usar la terminología exacta, lo cual se demuestra mediante métricas de calidad de texto como BLEU y ROUGE-2, que indican una mayor coincidencia terminológica y de frases cortas con el texto de referencia.

5.0.2. Análisis de errores

El análisis cualitativo revela tres focos de error:

- **Siglas y abreviaturas poco frecuentes.** Acrónimos idiosincrásicos o locales

pueden mantenerse sin expandirse o expandirse de forma no estándar si no están presentes en el corpus de ajuste.

- **Números y unidades.** En ejemplos con cuantificaciones escasas, aparecen desajustes tipográficos (espacio fino, signos) o variaciones de estilo (*mm* frente a *mm.*). Estos casos impactan poco en ROUGE pero son relevantes para la legibilidad clínica.
- **Negación y matices clínicos.** En oraciones muy breves, el modelo puede preferir reformulaciones estilísticas. Aunque los parámetros de decodificación son conservadores, la presión por fluidez puede introducir aclaraciones ligeras sin alterar el contenido factual.

5.0.3. Amenazas a la validez

- **Métricas centradas en superficie.** BLEU y ROUGE miden qué tan parecidas son las palabras y el orden en que aparecen, pero no verifican si la información médica es correcta o coherente. Para evaluación robusta proponemos incorporar CAS (Clinical Acceptability Score) y RadGraph-F1, que captan relaciones entre entidades y hallazgos.
- **Sesgo de dominio.** El corpus procede de un único entorno hospitalario; por tanto, estilos de dictado y convenciones locales pueden haber sido sobre-representados. La generalización a subdominios poco vistos (p. ej., pediatría) podría requerir *few-shot* adicional o ajuste por especialidad.
- **Tamaño del conjunto de prueba.** Cien informes ofrecen señal suficiente para tendencias globales, pero intervalos de confianza más ajustados requerirían conjuntos mayores y análisis por categoría (dependiendo del caso).

5.0.4. Implicaciones prácticas

- **Reducción de post-edición.** El aumento en BLEU/ROUGE se traduce operativamente en menos correcciones manuales de acentos, grafías y puntuación, y en frases clínicas más naturales (bigramas estables), con potencial ahorro de tiempo.
- **Consistencia terminológica.** La normalización de terminos compuestos como (*pósteroanterior*, *reticulonodular*) y de conectores mejora la uniformidad interinforme, facilitando auditorías y búsquedas.
- **Despliegue eficiente.** La combinación LoRA–QLoRA permite servir el modelo en hardware moderado manteniendo calidad, lo que favorece su integración.

6 Conclusiones y recomendaciones

6.1. Conclusiones

El proyecto cumplió el objetivo general de construir un sistema voz-a-texto orientado a informes radiológicos, con una arquitectura modular que encadena adquisición y normalización de audio, transcripción automática, corrección léxica y revisión semántica, y culmina en una interfaz de usuario capaz de entregar un informe listo para ser copiado o guardado. En términos de calidad, el ajuste fino específico del modelo Gemma-2B-IT bajo la estrategia LoRA sobre pesos cuantizados en 4 bits (QLoRA) demostró ser una vía eficaz y eficiente para especializar un modelo relativamente pequeño sin sacrificar estabilidad ni requerir infraestructura costosa: el sistema final superó al modelo base con un BLEU de 50.02 frente a 23.60 y mejoras notables en ROUGE-1, ROUGE-2 y ROUGE-L, lo que, en conjunto, se traduce en mayor precisión ortográfica y terminológica, mejor fluidez local de bigramas clínicos y mejor preservación de la estructura típica del reporte radiológico. Estas ganancias son coherentes con el diseño técnico del flujo: la normalización acústica reduce variabilidad de entrada; la decodificación con marcas temporales aporta trazabilidad; la corrección léxica basada en un modelo biomédico de lenguaje enmascarado y un diccionario médico dinámico corrige términos especializados y estandariza grafías; y la revisión semántica instruida guía el tono y la puntuación sin inventar hallazgos, respetando el registro clínico. La ejecución de manera local posible por la cuantización de pesos y la exportación a formatos eficientes, permite preservar la confidencialidad de los datos y reduce la dependencia de servicios externos, una consideración clave para hospitales con restricciones de conectividad o políticas estrictas de privacidad. Finalmente, la GUI integró controles de adquisición, estado del sistema y acciones de entrega del informe en una sola ventana, disminuyendo la carga cognitiva y favoreciendo la adopción en la práctica; en conjunto, la evidencia cuantitativa y cualitativa sugiere que un enfoque de especialización ligera, apoyado en recursos abiertos y gobernanza terminológica, puede ofrecer un desempeño modular y sostenible en entornos clínicos de la región.

6.2. Recomendaciones

En el ambito de la adopción institucional y en cuanto a la evolución técnica del sistema, se recomienda implementar una validación humana obligatoria durante las primeras fa-

ses de uso para mitigar riesgos clínicos y generar trazas de decisiones; mantener un ciclo de mejora continua del diccionario médico y de los prompts de revisión semántica mediante un bucle de retroalimentación donde los radiólogos acepten o corrijan sugerencias y estas decisiones alimenten iteraciones de ajuste; complementar BLEU/ROUGE con métricas clínicas orientadas a factualidad y relaciones entre entidades (p. ej., CAS o RadGraph-F1) y con revisiones ciegas por pares; estudiar el ajuste fino específico del reconocedor de voz para el dominio radiológico en español a fin de reducir la tasa de error de palabra antes de la corrección textual; explorar técnicas adicionales de eficiencia —entrenamiento consciente de cuantización, poda estructurada o destilación— para llevar la inferencia a GPU modestas o incluso a CPU manteniendo latencia aceptable; formalizar una política de gobernanza de datos que incluya anonimización robusta, control de acceso, auditoría y cumplimiento normativo local; establecer un plan de capacitación continua para personal clínico y de TI que cubra límites del sistema, manejo de excepciones y mejores prácticas de edición; instrumentar monitoreo operativo con alarmas sobre deriva de distribución, degradación de métricas o fallos de hardware, acompañado de procedimientos de revertimiento y de re-entrenamiento programado; y, por último, ampliar el corpus con datos de múltiples centros y subespecialidades para mejorar la generalización, realizando estudios de ablación que cuantifiquen el aporte individual de cada componente (normalización acústica, umbrales de confianza, rango LoRA, parámetros de decodificación) y ajustando el sistema a objetivos específicos de cada servicio (informe breve telegráfico frente a narrativo, por ejemplo), todo ello sin perder de vista que la interoperabilidad y la trazabilidad deben permanecer como principios rectores del diseño.

Bibliografía

Alharbi, S., Alrazgan, M., Alrased, A., Alnomasi, T., Almogel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almogil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. *IEEE Access*, *9*, 1-20. https://doi.org/10.1109/ACCESS.2021.3112535

Alqahtani, F. F., Mohsan, M. M., Alshamrani, K., Zeb, J., Alhamami, S., & Alqarni, D. (2024). CNX-B2: A Novel CNN-Transformer Approach For Chest X-Ray Medical Report Generation. *IEEE Access*, *12*, 26626-26635. https://doi.org/10.1109/ACCESS.2024.3367360

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E. K., Molaei, A., Jia, Y., Jose, A., Roy, R., & Merhof, D. (2024). Advances in medical image analysis with vision Transformers: A comprehensive review. *Medical Image Analysis*, *91*, 103000. https://doi.org/10.1016/J.MEDIA.2023.103000

Babar, Z., van Laarhoven, T., Zanzotto, F. M., & Marchiori, E. (2021). Evaluating diagnostic content of AI-generated radiology reports of chest X-rays. *Artificial Intelligence in Medicine*, *116*, 102075. https://doi.org/10.1016/j.artmed.2021.102075

Bitterman, D. S., Miller, T. A., Mak, R. H., & Savova, G. K. (2021). Clinical Natural Language Processing for Radiation Oncology: A Review and Practical Primer. *International Journal of Radiation Oncology*Biology*Physics*, *110*, 641-655. https://doi.org/10.1016/J.IJROBP.2021.01.044

Busch, F., Hoffmann, L., dos Santos, D. P., Makowski, M. R., Saba, L., Prucker, P., Hadamitzky, M., Navab, N., Kather, J. N., Truhn, D., Cuocolo, R., Adams, L. C., & Bressem, K. K. (2024). Large language models for structured reporting in radiology: past, present, and future. *European Radiology*, *35*, 2589-2602. https://doi.org/10.1007/S00330-024-11107-6/FIGURES/2

Busch, F., Prucker, P., Komenda, A., Ziegelmayer, S., Makowski, M. R., Bressem, K. K., & Adams, L. C. (2025). Multilingual feasibility of GPT-4o for automated Voice-to-Text CT and MRI report transcription. *European Journal of Radiology*, *182*, 111827. https://doi.org/10.1016/J.EJRAD.2024.111827

Carrino, C. P., Llop, J., Pàmies, M., Gutiérrez-Fandiño, A., Armengol-Estapé, J., Silveira-Ocampo, J., Valencia, A., Gonzalez-Agirre, A., & Villegas, M. (2022, mayo). Pretrained Biomedical Language Models for Clinical NLP in Spanish. En D. Demner-Fushman, K. B. Cohen, S. Ananiadou & J. Tsujii (Eds.), *Proceedings of the 21st Workshop on Biomedical Language Processing* (pp. 193-199). Associa-

tion for Computational Linguistics. https://doi.org/10.18653/v1/2022.bionlp-1.19

Casey, A., Davidson, E., Poon, M., Dong, H., Duma, D., Grivas, A., Grover, C., Suárez-Paniagua, V., Tobin, R., Whiteley, W., Wu, H., & Alex, B. (2021). A systematic review of natural language processing applied to radiology reports. *BMC Medical Informatics and Decision Making*, *21*(1), 179. https://doi.org/10.1186/s12911-021-01533-7

Chew, B. H., & Ngiam, K. Y. (2025). Artificial intelligence tool development: what clinicians need to know? *BMC Medicine*, *23*, 1-19. https://doi.org/10.1186/S12916-025-04076-0/TABLES/7

Consortium, U. (2025). Unicode Standard Annex #15: Unicode Normalization Forms. *Unicode Standard*. https://unicode.org/reports/tr15/

Corporation, M. (2025). Resource Interchange File Format (RIFF) Specification. *Microsoft Documentation*. https://learn.microsoft.com/en-us/windows/win32/multimedia/resource-interchange-file-format-riff

Czum, J. M. (2020). Dive Into Deep Learning. *Journal of the American College of Radiology*, *17*, 637-638. https://doi.org/10.1016/J.JACR.2020.02.005

Delbrouck, J.-B., Xu, J., Moll, J., Thomas, A., Chen, Z., Ostmeier, S., et al. (2025). Automated Structured Radiology Report Generation. *arXiv preprint arXiv:2505.24223*. https://doi.org/https://doi.org/10.48550/arXiv.2409.16563

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv*. https://arxiv.org/abs/2305.14314

Developers, N. (2025). NumPy — the fundamental package for array computing in Python. *NumPy Documentation*. https://numpy.org/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

Dubinski, D., Won, S. Y., Trnovec, S., Behmanesh, B., Baumgarten, P., Dinc, N., Konczalla, J., Chan, A., Bernstock, J. D., Freiman, T. M., & Gessler, F. (2024). Leveraging artificial intelligence in neurosurgery—unveiling ChatGPT for neurosurgical discharge summaries and operative reports. *Acta Neurochirurgica*, *166*, 1-6. https://doi.org/10.1007/S00701-024-05908-3/FIGURES/3

Falcetta, F. S., de Almeida, F. K., Lemos, J. C. S., Goldim, J. R., & da Costa, C. A. (2023). Automatic documentation of professional health interactions: A systematic review. *Artificial Intelligence in Medicine*, *137*, 102487. https://doi.org/10.1016/J.ARTMED.2023.102487

Fanni, S. C., Tumminello, L., Formica, V., Caputo, F. P., Aghakhanyan, G., Ambrosini, I., Francischello, R., Faggioni, · . L., Cioni, D., & Neri, · . E. (2024). The journey from natural language processing to large language models: key insights for radiologists. *Journal of Medical Imaging and Interventional Radiology 2024 11:1*, *11*, 1-10. https://doi.org/10.1007/S44326-024-00043-W

Foundation, P. S. (2025a). subprocess — Subprocess management. *Python3 Standard Library Documentation*. https://docs.python.org/3/library/subprocess.html

Foundation, P. S. (2025b). wave — Read and write WAV files. *Python3 Standard Library Documentation*. https://docs.python.org/3/library/wave.html

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., ... Ma, Z. (2024). The Llama 3 Herd of Models. https://arxiv.org/abs/2407.21783

Groot, O. Q., Ogink, P. T., Oosterhoff, J. H., & Beam, A. L. (2021). Natural language processing and its role in spine surgery: A narrative review of potentials and challenges. *Seminars in Spine Surgery*, *33*, 100877. https://doi.org/10.1016/J.SEMSS.2021.100877

Guo, K., Zheng, S., Huang, R., & Gao, R. (2023). Multi-Task Learning for Lung Disease Classification and Report Generation via Prior Graph Structure and Contrastive Learning. *IEEE Access*, *11*, 110888-110898. https://doi.org/10.1109/ACCESS.2023.3322425

Hans, P., et al. (2025). python-sounddevice — Play and Record Sound with Python. *Documentation, python-sounddevice*. https://python-sounddevice.readthedocs.io/

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., & Shen, D. (2023). Transformers in medical image analysis. *Intelligent Medicine*, *3*, 59-78. https://doi.org/10.1016/J.IMED.2022.07.002

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *arXiv*. https://arxiv.org/abs/2106.09685

Iqbal, S., Qureshi, A. N., Khan, F., Aurangzeb, K., & Akbar, M. A. (2024). From Data to Diagnosis: Enhancing Radiology Reporting With Clinical Features Encoding and Cross-Modal Coherence. *IEEE Access*, *12*, 127341-127356. https://doi.org/10.1109/ACCESS.2024.3449929

Jelassi, M., Jemai, O., & Demongeot, J. (2024). Revolutionizing Radiological Analysis: The Future of French Language Automatic Speech Recognition in Healthcare. *Diagnostics*, *14*(9), 895. https://doi.org/10.3390/diagnostics14090895

Jorg, T., Halfmann, M. C., Stoehr, F., Arnhold, G., Theobald, A., Mildenberger, P., & Müller, L. (2024). A novel reporting workflow for automated integration of artificial intelligence results into structured radiology reports. *Insights into Imaging*, *15*(80), 1-10. https://doi.org/10.1186/s13244-024-01660-5

Jorg, T., Kämpgen, B., Feiler, D., Müller, L., Düber, C., Mildenberger, P., & Jungmann, F. (2023). Efficient structured reporting in radiology using an intelligent dialogue system based on speech recognition and natural language processing. *Insights into Imaging*, *14*(47), 1-9. https://doi.org/10.1186/s13244-023-01392-y

Krishna, K., Khosla, S., Bigham, J. P., & Lipton, Z. C. (2021). Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques. https://arxiv.org/abs/2005.01795

Kudo, T. (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. https://arxiv.org/abs/1804.10959

Li, S., Qiao, P., Wang, L., Ning, M., Yuan, L., Zheng, Y., & Chen, J. (2024). An Organ-Aware Diagnosis Framework for Radiology Report Generation. *IEEE Transactions on Medical Imaging*, *43*, 4253-4265. https://doi.org/10.1109/TMI.2024.3421599

Lin, C., & Kuo, C.-F. (2025). Roles and Potential of Large Language Models in Healthcare: A Comprehensive Review. *Biomedical Journal*, 100868. https://doi.org/10.1016/J.BJ.2025.100868

Luo, J. W., & Chong, J. J. (2020). Review of Natural Language Processing in Radiology. *Neuroimaging Clinics of North America*, *30*, 447-458. https://doi.org/10.1016/J.NIC.2020.08.001

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781

Moezzi, S. A. R., Ghaedi, A., Rahmanian, M., Mousavi, S. Z., & Sami, A. (2023). Application of Deep Learning in Generating Structured Radiology Reports: A Transformer-Based Technique. *Journal of Digital Imaging*, *36*, 80-90. https://doi.org/10.1007/s10278-022-00692-x

NVIDIA. (2020). How to Build Domain-Specific Automatic Speech Recognition Models on GPUs [Accedido el 25 de mayo de 2025]. https://developer.nvidia.com/blog/how-to-build-domain-specific-automatic-speech-recognition-models-on-gpus/

Olex, A. L., & McInnes, B. T. (2021). Review of Temporal Reasoning in the Clinical Domain for Timeline Extraction: Where we are and where we need to be. *Journal of Biomedical Informatics*, *118*, 103784. https://doi.org/10.1016/J.JBI.2021.103784

Ozsahin, D. U., Usanase, N., & Ozsahin, I. (2025). Advancing pancreatic cancer management: the role of artificial intelligence in diagnosis and therapy. *Beni-Suef University Journal of Basic and Applied Sciences 2025 14:1*, *14*, 1-18. https://doi.org/10.1186/S43088-025-00610-4

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. https://arxiv.org/abs/1406.2038

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. https://arxiv.org/abs/1802.05365

Post, M. (2018). sacreBLEU — A hassle-free implementation of BLEU for reproducible machine translation evaluation. *GitHub Repository, sacreBLEU*. https://github.com/mjpost/sacrebleu/

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. https://arxiv.org/abs/2212.04356

Research/opensource, G. (2020). rouge-score — A native Python implementation of ROUGE metrics. *PyPI Package, rouge-score*. https://pypi.org/project/rouge-score/

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. https://arxiv.org/abs/1508.07909

Sheikhy, A., Fatemeh, · ., Firouzabadi, D., Lay, N., Jarrah, N., Pouria, · ., Anari, Y., & Malayeri, A. (2025). State of the art review of AI in renal imaging. *Abdominal Radiology 2025*, 1-19. https://doi.org/10.1007/S00261-025-04963-3

Taylor, A. M. (2022). The role of artificial intelligence in paediatric cardiovascular magnetism resonance imaging. *Pediatric Radiology*, *52*, 2131-2138. https://doi.org/10.1007/S00247-021-05218-1/FIGURES/3

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., ... Andreev, A. (2024). Gemma 2: Improving Open Language Models at a Practical Size. https://arxiv.org/abs/2408.00118

Tsaniya, H., Fatichah, C., & Suciati, N. (2024). Automatic Radiology Report Generator Using Transformer With Contrast-Based Image Enhancement. *IEEE Access*, *12*, 25429-25442. https://doi.org/10.1109/ACCESS.2024.3364373

Wang, D., & Zhang, S. (2024). Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial Intelligence Review*, *57*, 1-48. https://doi.org/10.1007/S10462-024-10921-0/TABLES/8

Wei, Y., Wang, X., Ong, H., Zhou, Y., Flanders, A., Shih, G., & Peng, Y. (2024). Enhancing disease detection in radiology reports through fine-tuning lightweight LLM on weak labels. https://arxiv.org/abs/2409.16563

Yang, Z., Wang, D., Zhou, F., Song, D., Zhang, Y., Jiang, J., Kong, K., Liu, X., Qiao, Y., Chang, R. T., Han, Y., Li, F., Tham, C. C., & Zhang, X. (2024). Understanding natural language: Potential application of large language models to ophthalmology. *Asia-Pacific Journal of Ophthalmology*, *13*, 100085. https://doi.org/10.1016/J.APJO.2024.100085

Yenduri, G., M, R., G, C. S., Y, S., Srivastava, G., Maddikunta, P. K. R., G, D. R., Jhaveri, R. H., B, P., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2023). Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. https://arxiv.org/abs/2305.10435

Zhang, Y., Pan, Y., Zhong, T., Dong, P., Xie, K., Liu, Y., Jiang, H., Liu, Z., Zhao, S., Zhang, T., Jiang, X., Shen, D., Liu, T., & Zhang, X. (2024). Potential of multimodal large language models for data mining of medical images and free-

text reports. *Meta-Radiology*, *2*, 100103. https://doi.org/10.1016/J.METRAD.2024.100103