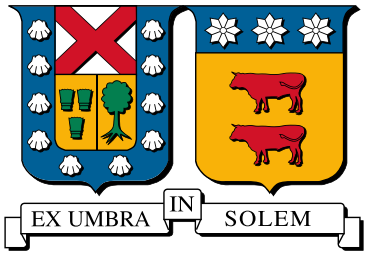


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

Departamento de Electrónica

Valparaíso - Chile

"AJUSTE FINO DE PARÁMETROS DE UN MODELO LLM MULTIMODAL PARA ONCOLOGÍA DE PRECISIÓN"

SEBASTIÁN GONZALO GAETE CAROCA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELECTRÓNICO.

PROFESOR GUÍA: DR. WERNER CREIXELL FUENTES
PROFESOR CORREFERENTE: DR. ALEJANDRO WEINSTEIN OPPENHEIMER

ENERO 2026

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☑ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Ajuste Fino De Parámetros de un Modelo LLM Multimodal para Oncología de Precisión

Nombre del candidato(a): Sebastián Gonzalo Gaete Caroca

Carrera / Grado: Ingeniería Civil Electrónica

Campus: Casa Central **Departamento:** Electrónica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Werner Creixell Fuentes, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☑ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

_____

_____

_____

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 20 de Abril 2026 **Firma:** _____

Estudiante o Candidato(a):

Fecha: 20 de abril 2026 **Firma:** _____

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

DEPARTAMENTO DE ELECTRONICA
UNIVERSIDAD TECNICA FEDERICO SANTA MARIA

DEPARTAMENTO DE ELECTRONICA

Ajuste Fino de Parámetros de un Modelo LLM Multimodal para Oncología de Precisión

Sebastián Gonzalo Gaete Caroca

Memoria para optar al título de Ingeniero Civil Electrónico, mención Computadores, submención Control e Instrumentación.

Universidad Técnica Federico Santa María

Profesor Guía: Dr. Werner Creixell Fuentes

Profesor Correferente: Dr. Alejandro Weinstein Oppenheimer

ENERO 2026

Este trabajo presenta la adaptación y *fine-tuning* de un modelo LLM multimodal preentrenado aplicado en el contexto de oncología de precisión. El objetivo principal es generar automáticamente informes microscópicos a partir de *Whole Slide Images (WSI)* de cáncer de mama, proporcionando una herramienta de asistencia para el patólogo que optimice el flujo de trabajo y reduzca la carga cognitiva de una tarea diagnóstica crítica. El documento detalla desde la revisión del estado del arte de los modelos multimodales, tanto de propósito general como específicos para el dominio medico, hasta la implementación técnica, la validación del modelo generado y la evaluación de sus resultados. Finalmente se discuten las conclusiones sobre el diseño y el desempeño del modelo en comparación con los estándares actuales y benchmarks relevantes.

DEPARTAMENTO DE ELECTRONICA
UNIVERSIDAD TECNICA FEDERICO SANTA MARIA

DEPARTAMENTO DE ELECTRONICA

Fine-Tuning of Parameters of a Multimodal LLM for Precision Oncology

Sebastián Gonzalo Gaete Caroca

Thesis for the fulfillment of the B.S. in Electronic Engineering, major in Computer Electronics, minor in Control and Instrumentation (6 year program).

Universidad Técnica Federico Santa María

Advisor: Dr. Werner Creixell Fuentes

Co-Advisor: Dr. Alejandro Weinstein Oppenheimer

ENERO 2026

This work presents the adaptation and fine-tuning of a pre-trained multimodal large language model (LLM) applied in the context of precision oncology. The primary objective is to automatically generate microscopic reports from breast cancer *Whole Slide Images (WSIs)*, providing an assistive tool for pathologists that optimizes workflow and reduces the cognitive burden associated with a critical diagnostic task. The document details the process from a review of the state of the art in multimodal models-both general-purpose and domain-specific to the medical field-through to the technical implementation, model validation, and evaluation of the obtained results. Finally, conclusions are discussed regarding the model's design and performance in comparison with current standards and relevant benchmarks.

Glosario

Benchmark Conjunto estandarizado de tareas, métricas y datasets utilizado para evaluar y comparar el desempeño de modelos computacionales.

Dataset Conjunto estructurado de datos utilizado para entrenamiento, validación o evaluación de modelos de aprendizaje automático.

Fine-tuning Proceso de ajuste adicional de un modelo preentrenado utilizando datos específicos de una tarea particular para especializar su desempeño.

H&E Hematoxilina y Eosina; tinción histológica estándar que resalta núcleos celulares en azul y citoplasma y matriz extracelular en tonos rosados.

Histopatología Disciplina médica que estudia las alteraciones microscópicas de los tejidos para el diagnóstico y caracterización de enfermedades.

LLM Large Language Model; modelo de lenguaje basado en arquitecturas profundas, entrenado con grandes volúmenes de texto para tareas de generación y comprensión.

Patólogo Especialista médico encargado del análisis microscópico de tejidos para establecer diagnósticos y clasificaciones histológicas.

TCGA The Cancer Genome Atlas; proyecto internacional que proporciona datos genómicos y clínicos de múltiples tipos de cáncer para investigación biomédica.

Transformer Arquitectura de redes neuronales basada en mecanismos de atención, ampliamente utilizada en modelos de lenguaje y sistemas multimodales.

VQA Visual Question Answering; tarea multimodal que consiste en responder preguntas en lenguaje natural a partir del contenido de una imagen.

WSI Whole Slide Image; imagen digital de alta resolución que representa una lámina histológica completa escaneada a nivel microscópico.

Índice de contenidos

Índice de figuras

CAPÍTULO 1

Introducción

La oncología es una de las áreas más relevantes de la medicina moderna, ya que el cáncer continúa siendo una de las principales causas de mortalidad a nivel mundial [1] [2]. Tradicionalmente, los tumores se han clasificado principalmente según el órgano de origen. Sin embargo, los avances en biología molecular y patología han permitido caracterizar el cáncer en función de alteraciones genéticas, moleculares y morfológicas específicas. Este enfoque dió origen a la oncología de precisión, cuyo objetivo es adaptar el diagnóstico y el tratamiento a las características particulares de cada paciente y de su tumor [3]. Este enfoque permite apoyar decisiones terapéuticas personalizadas, considerando el tipo de tratamiento, la dosis y el momento de administración. A pesar de estos avances, persisten desafíos importantes. La confidencialidad de los datos, la complejidad del análisis y la subjetividad diagnóstica dificultan su implementación a gran escala.

En este contexto, los modelos de inteligencia artificial surgen como herramientas de apoyo. En particular, los modelos fundacionales de *Deep Learning* en visión y lenguaje han demostrado alta capacidad de análisis en distintos dominios [4]. Desde 2023 [5], se ha incrementado el desarrollo de modelos y herramientas capaces de procesar *Whole Slide Images (WSI)* para la detección y caracterización de tumores. Estas imágenes pueden alcanzar dimensiones del orden de $100\,000 \times 100\,000$ píxeles. Debido a esta complejidad, la generación automática de reportes a partir de WSI sigue siendo un desafío abierto y una etapa crítica dentro del flujo de trabajo del patólogo. En consecuencia, este trabajo busca desarrollar una herramienta de asistencia para dicha tarea. Para abordar este desafío se acota el dominio del problema centrándose exclusivamente en cáncer de

mama. Esta elección se justifica por su alta incidencia a nivel mundial y por la mayor disponibilidad de datos en comparación con otros tipos de cáncer [6] [7]. El desempeño del sistema propuesto se evalúa mediante métricas tradicionales de procesamiento de lenguaje natural, junto con métricas específicas orientadas a capturar la precisión clínica y la relevancia patológica de las respuestas generadas.

1.1. Propuesta de solución

Para abordar el problema planteado, se propone un sistema basado en un modelo de visión y un modelo de lenguaje. El modelo de visión se encarga de extraer las características relevantes de la *Whole Slide Image (WSI)*. A partir de esta información, el modelo de lenguaje (LLM) genera un reporte estructurado y coherente.

El sistema recibe como entrada una *WSI* junto con una consulta y produce como salida un reporte o respuesta asociada a dicha imagen. El desempeño del sistema depende de factores clave, como la disponibilidad de datos, la capacidad de almacenamiento y los recursos de cómputo. Estos factores influyen directamente en la calidad del reporte generado. Por esta razón, se busca maximizar la calidad de la respuesta utilizando los recursos disponibles.

1.2. Estructura del documento

En el siguiente capítulo, *Estado del Arte*, se revisan investigaciones relevantes del área, incluyendo técnicas de procesamiento y entrenamiento, datasets disponibles, modelos actuales, soluciones existentes y benchmarks utilizados.

El capítulo 3, *Diseño e Implementación de la Solución*, describe la arquitectura propuesta, la estrategia de entrenamiento, el dataset empleado y el hardware disponible.

En el capítulo 4, *Resultados*, se evalúa la calidad de la respuesta del modelo utilizando los benchmarks empleados en el área.

Finalmente, en el capítulo 5, *Conclusiones*, se presenta una síntesis de los resultados obtenidos y se proponen posibles líneas de trabajo futuro.

CAPÍTULO 2

Estado del Arte

2.1. Histopatología digital y flujo de trabajo

El flujo de trabajo tradicional en histopatología, llevada a cabo por un patólogo, era (y en general sigue siendo) fundamentalmente manual. Las muestras de tejido se colocan en portaobjetos de vidrio y son examinadas mediante microscópicos ópticos convencionales [8]. El análisis depende exclusivamente de la evaluación visual directa. Esto lo hace un proceso lento y propenso a sesgos.

La transición hacia lo digital comenzó introduciendo la *Whole Slide Imaging, WSI*. Esto permite escanear el portaobjetos de vidrio y generar imágenes digitales de alta resolución, habilitando su visualización y análisis en entornos computacionales [9]. Un ejemplo de WSI se puede ver en la Figura 1.1. La digitalización completa de láminas con Whole Slide Imaging (WSI) ha sido evaluada y adoptada para diagnóstico patológico primario en contextos clínicos regulados [10] [11].

Paralelamente surge la *Computational Pathology* (CPATH), definido como un enfoque de análisis basado en grandes volúmenes de datos. Integra imágenes histopatológicas y metadatos clínicos para extraer características y ser usados, por ejemplo, con técnicas de inteligencia artificial [12].

En la práctica actual, el análisis de *Whole Slide Images* se realiza mediante visores digitales como *QuPath* [13], que permiten desplazamiento y múltiples niveles de aumento.

Las aplicaciones de la histopatología computacional abarca tanto en el ámbito clínico

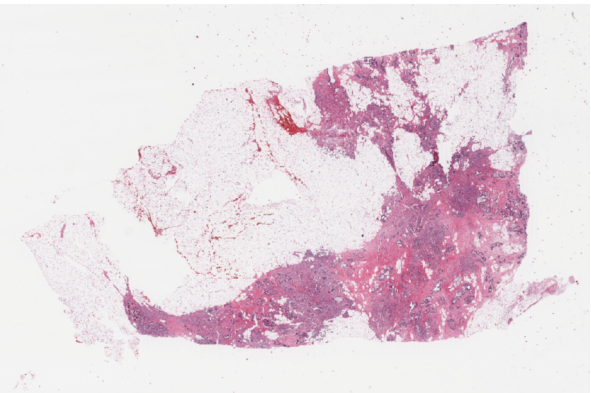


Figura 1.1: WSI TCGA-BH-A0DP-01Z-00-DX1 de TCGA

como en el de investigación. Entre ellas se incluyen herramientas de asistencia diagnostica, como la detección de metástasis, conteo de mitosis o clasificación tumoral [5].

La implementación de estas tecnologías requiere una infraestructura computacional adecuada. El entrenamiento de modelos de *Deep Learning* se beneficia del uso de GPUs. Mientras que el almacenamiento y transmisión de datos representan un desafío, debido al gran tamaño de las WSI. Una WSI puede variar en tamaño de 300MB a 4GB aproximadamente.

Finalmente, existe un esfuerzo creciente por mejorar la interpretabilidad de los modelos, de tener sistemas mas transparentes, con el objetivo de aumentar la confianza del patólogo en estas herramientas.

2.2. Procesamiento de Whole Slide Image

El procesamiento de *Whole Slide Images* (WSI) requiere herramientas especializadas debido a su gran tamaño, el cual suele exceder la capacidad de memoria RAM y puede alcanzar varios gigabytes. Comúnmente, estas imágenes se almacenan en formatos piramidales, como `.svs` y `.tif`, que permiten múltiples niveles de resolución.

OpenSlide [14] es una biblioteca ampliamente utilizada para la lectura y manejo de WSI en distintos formatos. Proporciona una interfaz para acceder de forma eficiente a regiones específicas de la imagen y a diferentes niveles de magnificación, sin necesidad de cargar la imagen completa en memoria. Debido a estas características, OpenSlide se ha convertido en una herramienta base en numerosos sistemas de histopatología digital.

CLAM (*A Deep Learning-based Pipeline for Data Efficient and Weakly Supervised Whole-Slide-level Analysis*) [15] es un framework de código abierto para el análisis de WSI a nivel global. CLAM se apoya en OpenSlide para la lectura de imágenes y propone un flujo de procesamiento que incluye segmentación de tejido, extracción de parches y clasificación a nivel de imagen completa.

La segmentación del tejido permite descartar regiones de fondo irrelevantes, mientras que el parchado divide la WSI en imágenes de menor tamaño, típicamente del orden de 256×256 o 512×512 píxeles, compatibles con los modelos de visión. Este enfoque es consistente con la literatura, dado que la mayoría de los modelos de visión operan sobre imágenes de tamaño fijo y resolución limitada.

Además, CLAM incorpora un módulo de *feature extraction*, en el cual los parches son transformados en representaciones vectoriales (*embeddings*) mediante modelos de visión adaptados al dominio de la histopatología. Estas representaciones permiten realizar análisis posteriores a nivel de portaobjetos utilizando esquemas de aprendizaje débilmente supervisado.

2.3. Modelos de Visión

Las redes convolucionales (CNN) han demostrado un alto desempeño en el análisis de [16]. Sin embargo, la introducción de arquitecturas basadas en *Transformers* [17] permitió mejorar la extracción de características al incorporar contexto global mediante mecanismos de atención.

En modelos de visión basados en *Transformers*, como se muestra en la Figura 3.2 las imágenes se dividen en parches de tamaño fijo, típicamente 16×16 píxeles [18]. Cada parche se proyecta a un token visual que es procesado por capas de *Self-Attention*, generando representaciones enriquecidas con contexto espacial. Este enfoque ha mostrado ventajas frente a las CNN en tareas de descripción y comprensión visual de propósito general. En el dominio de la histopatología, estos avances han motivado el desarrollo de modelos de visión especializados.

CONCH (*Contrastive Learning from Captions for Histopathology*) [19] es un modelo fundacional de visión-lenguaje diseñado específicamente para histopatología. Está basado en el marco CoCa [20] e integra un codificador de imágenes, un codificador de texto

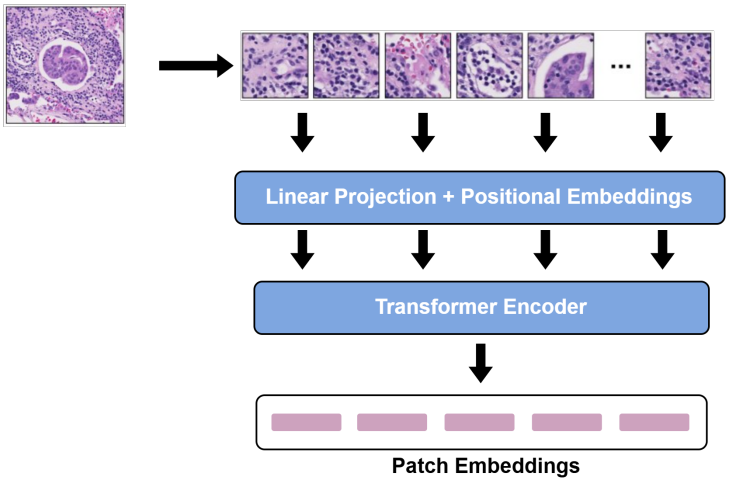


Figura 3.2: Arquitectura general del Vision Transformer

y un decodificador multimodal. El modelo se entrena mediante objetivos contrastivos y de generación de *captions*. El preentrenamiento de CONCH se realizó sobre más de 1.17 millones de pares imagen-texto, constituyendo uno de los mayores conjuntos multimodales en este dominio. CONCH supera modelos previos como PLIP y BiomedCLIP en diversos tipos de tareas. Además, destaca por su capacidad de transferencia *zero-shot* y recuperación intermodal imagen-texto.

UNI [21] es un modelo de visión auto-supervisado de propósito general basado en la arquitectura Vision Transformer Large (ViT-Large). A diferencia de los enfoques multimodales, UNI se centra en el aprendizaje de representaciones visuales mediante auto-supervisión. El entrenamiento se realiza utilizando el algoritmo DINOv2 sobre un dataset que contiene más de 100 millones de parches extraídos de más de 100.000 *Whole Slide Images* de 20 tipos de tejidos.

En evaluaciones sobre 33 tareas clínicas, UNI supera a codificadores de referencia como CTransPath y REMEDIS. El modelo demuestra una capacidad de generalización robusta en tareas de distinta complejidad diagnóstica, incluyendo la clasificación de hasta 108 tipos de cáncer.

En conjunto, CONCH y UNI representan un avance significativo en la literatura al abordar la escasez de anotaciones en histopatología mediante preentrenamiento a gran escala. Ambos modelos permiten flujos de trabajo que requieren un ajuste fino supervisado mínimo o nulo, facilitando su integración en sistemas posteriores de análisis y

generación de reportes.

2.4. Modelos de Lenguaje y MLLM

La generación de reportes en histopatología requiere modelos capaces de integrar información visual y textual. Para esta tarea, las arquitecturas multimodales mas utilizadas son: *decoder-only* y *cross-attention* [22].

En el enfoque *decoder-only* (Figura 4.3), las imágenes se procesan mediante un codificador de visión que divide la imagen en parches y genera representaciones vectoriales. Estas representaciones se proyectan al espacio del modelo de lenguaje y se concatenan directamente con los tokens de texto. La secuencia combinada es procesada por un único decodificador sin modificaciones en el LLM base. Esta forma destaca por su simplicidad de implementación y compatibilidad con LLMs preentrenados.

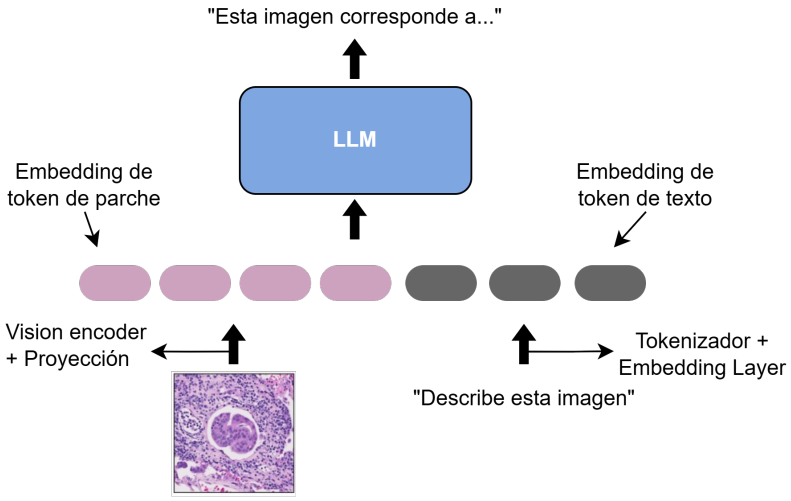


Figura 4.3: Enfoque decoder only en LLMs

El enfoque basado en *cross-attention* introduce mecanismos de atención cruzada para integrar información visual en las capas internas del transformador. En este caso las *querys* vienen del decodificador textual, mientras que las *keys* y *values* se derivan del codificador de imágenes. Esta arquitectura permite una integración mas profunda de ambas modalidades. Entre sus ventajas se encuentran una mayor eficiencia computacional y una mejor gestión de imágenes de alta resolución, al evitar la sobrecarga de la ventana de contexto.

El desarrollo de modelos multimodales se apoya en LLMs entrenados sobres datos masivos. Entre los modelos abiertos mas relevantes se encuentran Llama y Qwen.

LLama 3 [23] es una familia de modelos basados en la arquitectura Tranformer, diseñada para maximizar la estabilidad del entrenamiento. El modelo fue preentrenado con 15 billones de tokens multilingües y ajustado mediante técnicas como *supervised fine-tuning* y DPO. Soporta ventanas de contexto de hasta 128k tokens. Su variante multimodal, Llama 3.2, integra imágenes mediante un mecanismo de atención cruzada sin modificar los parámetros del modelo base.

Qwen 2.5 [24] es una familia de modelos que abarca versiones densas y de mezcla de expertos. Fue entrenado con 18 billones de tokens e incorpora técnicas avanzadas de fine-tuning y reinforcement learning. Destaca por su desempeño en matemáticas y programación, así como por su capacidad de manejar contextos extensos. Su variante multimodal introduce mecanismos de resolución dinámica para el procesamiento de imágenes.

Entre los modelos multimodales abiertos de propósito general destacan Gemma y LLaVA, los cuales han servido como base para adaptaciones en el dominio médico.

Gemma 3 [25] es una familia de modelos multimodales diseñados para ser eficientes en hardware de consumo. Integra un codificador de visión basado en SigLIP y representa las imágenes como secuencias compactas de tokens visuales. Soporta contextos largos mediante una combinación de atención local y global. Su entrenamiento incluye knowledge destilation desde modelos de mayor escala.

LLaVA (*Large Language and Vision Assistant*) [26] es un marco de trabajo centrado en la sintonización de instrucciones visuales. Su arquitectura conecta un codificador de visión preentrenado con un modelo de lenguaje mediante un proyector ligero. El entrenamiento se realiza en dos etapas: alineación de características visuales y sintonización de instrucciones multimodales. Este enfoque ha demostrado ser efectivo para extender las capacidades conversacionales de los LLMs hacia el dominio visual.

En histopatología, los modelos multimodales han sido adaptados para abordar desafíos específicos como la alta resolución de las imágenes, la necesidad de explicabilidad y la escasez de datos anotados.

Quilt-LLaVA [27] es un modelo multimodal para histopatología que describe regiones médicas relevantes en parches de WSIs y razona diagnósticos combinando evidencia

global. Se entrena con Quilt-Instruct, un gran dataset con anclaje espacial y razonamiento contextual, logrando mayor conciencia espacial. El dataset es generado a partir de videos educativos de histopatología.

WSI-LLaVA [28] está diseñado específicamente para el análisis de *Whole Slide Images*. El modelo prioriza la explicabilidad al generar descripciones morfológicas detalladas antes de emitir un diagnóstico. Para manejar imágenes de gigapíxeles, integra un codificador a nivel de parche basado en DINOv2 y un codificador a nivel de diapositiva basado en LongNet. El entrenamiento se realiza en tres etapas y se apoya en métricas específicas que evalúan la precisión clínica y la relevancia de las explicaciones generadas.

MedGemma [29] es una familia de modelos médicos fundacionales construidos sobre la arquitectura Gemma. Incorpora un codificador de visión especializado, MedSigLIP, entrenado con pares imagen-texto médicos y parches de histopatología. Gracias a este ajuste, MedGemma alcanza un rendimiento competitivo en tareas de histopatología, incluyendo clasificación de parches y análisis morfológico, manteniendo una buena capacidad de generalización.

2.5. Datasets de WSI, generación de reportes y VQA

El acceso a imágenes de *Whole Slide Images* en histopatología está fuertemente limitado por restricciones éticas y de confidencialidad, ya que los datos suelen estar asociados a información clínica sensible. En consecuencia, la disponibilidad de datasets públicos a gran escala para el entrenamiento de modelos multimodales en este dominio es reducida.

Uno de los recursos públicos más relevantes es *The Cancer Genome Atlas* (TCGA) [7], un programa de referencia que caracterizó molecularmente más de 20.000 tumores primarios correspondientes a 33 tipos de cáncer. A lo largo de más de una década, TCGA generó más de 2.5 petabytes de datos genómicos, epigenómicos, transcriptómicos, proteómicos y de imagen, los cuales se mantienen disponibles para la comunidad científica.

TCGA proporciona acceso a miles de WSIs, incluyendo un volumen significativo de casos de cáncer de mama. No obstante, dado que los reportes patológicos están asociados a la diapositiva completa, solo aquellos conjuntos con anotaciones a nivel WSI resultan adecuados para tareas de razonamiento multimodal y generación de lenguaje natural,

descartando datasets orientados exclusivamente a clasificación de parches o análisis puramente local.

Con el objetivo de abordar esta limitación, trabajos recientes han desarrollado datasets públicos orientados al *instruction tuning*, en los cuales cada WSI se asocia a pares de entrada-salida textuales y a representaciones embebidas de la diapositiva. Entre los esfuerzos más relevantes se encuentran SlideInstruction [30] y el set de entrenamiento de WSI-Bench [28], ambos construidos principalmente a partir de datos del TCGA, pero con enfoques metodológicos distintos.

SlideChat [30] introduce dos recursos complementarios: SlideInstruction, destinado al entrenamiento mediante seguimiento de instrucciones, y SlideBench, orientado a la evaluación exhaustiva de modelos multimodales en histopatología.

SlideInstruction [30] es presentado como uno de los mayores conjuntos de datos de seguimiento de instrucciones para WSIs disponibles al momento de su publicación. El dataset contiene 4.181 pares WSI-descripción y 175.753 pares de preguntas y respuestas visuales (VQA), derivados de informes de patología del TCGA.

La generación de los pares VQA se realizó mediante el uso de GPT-4, el cual fue empleado para limpiar los informes patológicos originales y generar preguntas abiertas y cerradas coherentes con el contenido clínico de cada caso.

Para la evaluación del modelo, los autores proponen SlideBench, un benchmark compuesto por tres subconjuntos diseñados para medir tanto el desempeño clínico como la capacidad de generalización.

WSI-LLaVA [28] propone WSI-Bench, un dataset unificado para entrenamiento y evaluación con un énfasis explícito en el análisis morfológico y la explicabilidad. Este conjunto se presenta como el primer benchmark a gran escala consciente de la morfología para WSIs de gigapíxeles.

En términos de escala y diversidad, WSI-Bench abarca 30 tipos de cáncer y contiene 9.850 WSIs correspondientes a 8.368 pacientes, superando a SlideChat en cobertura patológica. El dataset incluye un total de 179.569 pares VQA.

Las tareas están diseñadas para evaluar cuatro capacidades patológicas fundamentales a través de once tareas específicas: análisis morfológico, diagnóstico histopatológico, planificación del tratamiento y generación de informes clínicos completos.

2.6. Estrategias de entrenamiento

El entrenamiento de modelos LLM multimodales en histopatología pueden presentar altos requerimientos computacionales, especialmente al adaptar modelos de lenguaje preentrenados a dominios médicos específicos. En este contexto, se han propuesto distintas estrategias de ajuste fino que buscan equilibrar rendimiento y eficiencia.

El ajuste fino completo (*full fine-tuning*) [31] consiste en actualizar todos los parámetros del modelo preentrenado durante el entrenamiento. Este enfoque es considerado el estándar para maximizar el rendimiento cuando se adapta un modelo a una nueva tarea o dominio. Si bien suele ofrecer resultados superiores, su costo computacional es elevado. En modelos de miles de millones de parámetros, como LLaMA-7B, el ajuste completo requiere múltiples GPUs, estrategias de paralelización complejas y el almacenamiento de una copia completa de los pesos por tarea, lo que limita su viabilidad práctica.

Para mitigar estos costos, se han desarrollado técnicas de ajuste eficiente en parámetros [32] (*Parameter-Efficient Fine-Tuning*, PEFT), las cuales mantienen congelado el modelo base y entrenan únicamente un subconjunto reducido de parámetros adicionales.

Prefix Tuning [33] introduce tensores entrenables, denominados prefijos, en cada bloque del transformador. Estos prefijos modifican el comportamiento del modelo sin alterar los pesos originales. El método permite alcanzar un rendimiento comparable al ajuste completo entrenando aproximadamente el 0.1 % de los parámetros totales. Sin embargo, durante la inferencia es necesario suministrar explícitamente los prefijos aprendidos para activar la adaptación a la tarea.

El enfoque de adaptadores [31], inserta capas adicionales dentro de cada bloque del transformador, manteniendo congelado el resto del modelo. Estas capas siguen una arquitectura de cuello de botella, reduciendo significativamente el número de parámetros entrenables. Estudios previos muestran que modelos con adaptadores pueden igualar el rendimiento del ajuste completo entrenando menos del 5 % de los parámetros. No obstante, suelen requerir más parámetros que Prefix Tuning para alcanzar resultados similares.

LoRA [34] propone modelar la adaptación como una modificación de bajo rango de las matrices de pesos del modelo. En lugar de entrenar directamente los pesos originales,

estos se mantienen congelados y se introducen matrices de bajo rango entrenables cuya multiplicación representa el cambio en los pesos, esto se puede ver en la Figura 6.4. Este enfoque se basa en la hipótesis de que la adaptación a nuevas tareas reside en un subespacio de baja dimensión. LoRA ofrece una relación favorable entre rendimiento y eficiencia, y permite fusionar los parámetros aprendidos con el modelo base, evitando latencia adicional durante la inferencia.

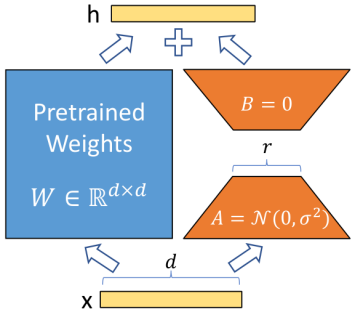


Figura 6.4: Reparametrización, A y B matrices entrenables

QLoRA [35] extiende LoRA mediante la cuantización del modelo base a 4 bits durante el entrenamiento. Esta estrategia reduce significativamente el consumo de memoria, permitiendo entrenar grandes modelos en hardware más limitado. Aunque el entrenamiento es más lento debido a la cuantización y decuantización, el impacto en la calidad final del modelo es mínimo en comparación con LoRA estándar, lo que ha favorecido su adopción en escenarios con restricciones de recursos.

En la literatura reciente, LoRA y QLoRA se consolidan como las técnicas PEFT más utilizadas para el entrenamiento. Estas estrategias permiten adaptar modelos fundacionales con un costo computacional reducido, manteniendo un rendimiento competitivo frente al ajuste fino completo.

2.7. Evaluación y Benchmarks

La evaluación de modelos multimodales generativos en histopatología presenta desafíos significativos, debido a la naturaleza abierta de las tareas de generación de texto y a la necesidad de capturar información clínicamente relevante más allá de la similitud lingüística superficial.

En la literatura, métricas clásicas de evaluación de lenguaje natural, como BLEU, ROUGE y METEOR, han sido ampliamente utilizadas para evaluar tareas de generación de texto abierto.

BLEU [36] mide el solapamiento n-gram entre la salida generada y la referencia, ROUGE [37] se centra en la cobertura de subsecuencias comunes, y METEOR [38] incorpora alineación semántica y penalización por fragmentación. Estas métricas evalúan principalmente similitud léxica y estructural. Sin embargo, en el contexto de histopatología, estas métricas resultan limitadas, ya que no evalúan la veracidad clínica ni la relevancia de los hallazgos patológicos descritos. Esto ocurre porque las métricas NLU tradicionales priorizan la similitud superficial del texto, sin distinguir entre afirmaciones clínicamente correctas y errores críticos con alto impacto diagnóstico. Por ejemplo, una respuesta que diagnostica un tumor como grado histológico 1 en lugar de grado histológico 3 puede obtener una alta puntuación, pese a diferir solo en un valor numérico, aun cuando ambas conclusiones implican comportamientos biológicos, pronóstico y decisiones terapéuticas completamente distintos.

Para abordar estas limitaciones, se ha propuesto el uso de modelos de lenguaje como jueces (*LLM-as-a-Judge*) [39], capaces de evaluar respuestas generadas considerando consistencia semántica, razonamiento clínico y coherencia contextual. Este enfoque permite una evaluación más alineada con criterios humanos, especialmente en tareas abiertas donde no existe una única respuesta correcta.

No obstante, la evaluación basada en LLMs introduce desafíos adicionales, como sesgos del modelo evaluador, falta de interpretabilidad de las puntuaciones y dependencia del modelo juez seleccionado, además de la calidad de la rúbrica de evaluación usada por el modelo.

Otro método, expuesto por WSI-LLaVA es evaluar la similitud de afirmaciones predichas del modelo y el ground truth [28]. El proceso consta de tres etapas: extracción de afirmaciones clínicas a partir del ground truth utilizando un LLM; evaluación individual de cada afirmación en la respuesta generada mediante un esquema de puntuación discreto; y cálculo del puntaje final como el promedio de las evaluaciones individuales. Las puntuaciones reflejan distintos niveles de corrección clínica, desde alineación completa con los hechos hasta información incorrecta o irrelevante, permitiendo una interpretación directa del desempeño del modelo.

En consecuencia, la tendencia actual en el estado del arte apunta a complementar

o reemplazar las métricas NLU clásicas por evaluaciones basadas en razonamiento, juicios automáticos asistidos por LLMs y métricas clínicas diseñadas específicamente para *WSIs*.

CAPÍTULO 3

Diseño e Implementación de la Solución

3.1. Limitaciones

El diseño e implementación del modelo estuvieron fuertemente condicionados por las restricciones de hardware disponibles. Las características del computador utilizado son las siguientes: CPU Intel Core i9-9980XE. 128GB DDR4 de memoria RAM. 10TB de almacenamiento secundario. 4 GPU TITAN RTX de 24GB cada una. Sistema operativo Ubuntu 20.04.6 LTS y versión 12.8 de CUDA.

Dado a que el diseño e implementación de un esquema de entrenamiento distribuido escapa al alcance del trabajo, el entrenamiento se realizó utilizando una única GPU de 24GB. Optando por priorizar estabilidad del entrenamiento y la reproducibilidad de los resultados, además de la utilización de una serie de técnicas para reducir los requerimientos de memoria y cómputo. Si bien se exploraron estrategias de paralelización, tales como *data parallelism* o *model parallelism* su aplicación resulta no trivial debido a los módulos y largas secuencias que implica un modelo de estas características.

3.2. Arquitectura

La arquitectura propuesta corresponde a un modelo multimodal visión-lenguaje diseñado para el análisis de WSI, combinando información visual a nivel de parches con capacidades generativas de lenguaje natural. El objetivo final es generar reportes o res-

puestas textuales coherentes a partir del contenido histopatológico de una WSI, bajo las restricciones computaciones descritas anteriormente. De manera general, la arquitectura se compone de tres módulos principales. Un encoder visual a nivel de parches, encargado de extraer representaciones visuales alineadas con el lenguaje. Un modelo de atención de largo alcance, que captura dependencias globales entre parches. Y finalmente, un modelo generativo, que produce la salida textual a partir de las representaciones visuales contextualizadas. El diagrama entrada/salida del sistema desarrollado se puede apreciar mejor en la Figura 2.1.

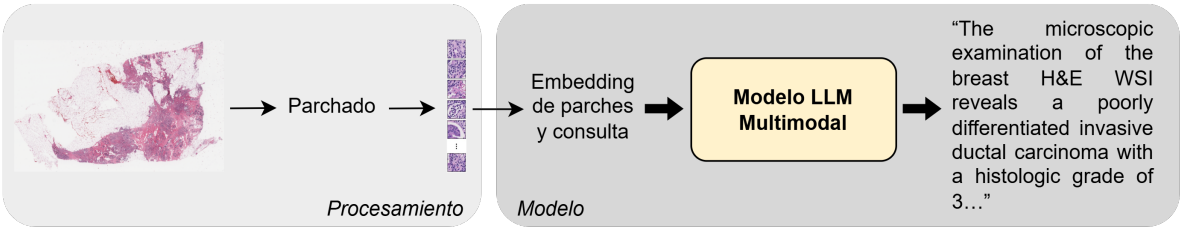


Figura 2.1: Flujo entrada/salida del sistema desarrollado

Dado el tamaño gigapíxel de las WSI, estas son divididas en parches utilizando CLAM, permitiendo un procesamiento a nivel local. Cada parche es posteriormente codificado mediante un *Vision Transformer* (ViT), actuando como encoder visual (ver Figura 3.2). Siguiendo el estado del arte en histopatología computacional, se utiliza CONCH como modelo visual base. CONCH es un ViT preentrenado en 1.17 millones de pares imagen-texto de histopatología. Su arquitectura base es de 12 capas Transformer, 12 cabezas de atención y tamaño de parche de 16×16. El modelo produce embeddings visuales de dimensión 512 por parche, las cuales se encuentran alineados semánticamente con representaciones textuales, facilitando la integracion multimodal posterior. Este módulo constituye el *patch-level encoder*, proporcionando representaciones locales ricas desde el punto de vista histopatológico, pero aún carentes de información contextual global.

Si bien los embeddings extraídos por el encoder visual contienen información relevante a nivel local, estos no incorporan relaciones entre parches, lo cual es crítico para la interpretación histopatológica a escala de WSI. Capturar patrones espaciales y dependencias globales requiere computar atención entre un número potencialmente muy elevado de parches, que puede superar los 10.000 e incluso alcanzar los 30.000 por WSI. El uso de mecanismos de atención estándar resulta computacionalmente inviable en este escenario, debido a su complejidad cuadrática con respecto a la longitud de

la secuencia. Para abordar este problema, se emplea LongNet [40] como módulo de agregación contextual. LongNet introduce el mecanismo de *Dilated Attention*, el cual permite modelar dependencias de largo alcance mediante un esquema de atención jerárquico y escalonado, reduciendo significativamente el costo computacional. En este enfoque, la atención se calcula a distintas escalas, comenzando con interacciones locales densas y progresivamente incorporando relaciones más distantes mediante saltos crecientes entre tokens. Este mecanismo permite capturar información global de la WSI sin necesidad de computar atención completa entre todos los parches. Una ilustración del funcionamiento de la atención dilatada se presenta en la Figura 2.2.

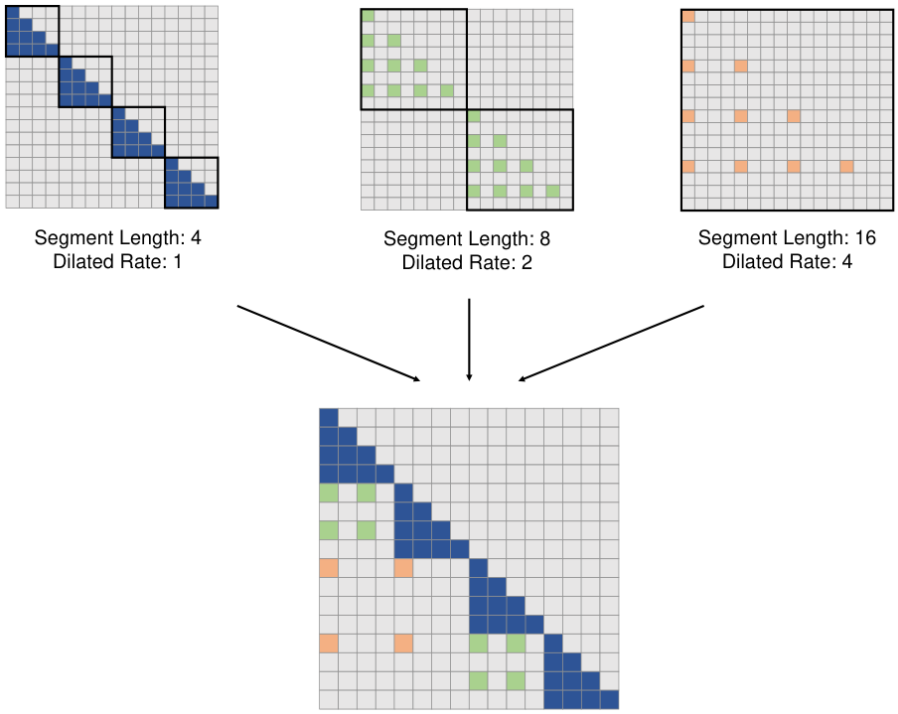


Figura 2.2: Esquema ilustrativo del mecanismo de Dilated Attention utilizado en LongNet.

La salida de este módulo corresponde a embeddings de parches enriquecidos con contexto global, manteniendo la dimensión original de 512.

Para permitir la integración de las representaciones visuales contextualizadas con el modelo generativo de lenguaje, es necesario alinear las dimensiones de los embeddings. Dado que el encoder visual produce vectores de dimensión 512, el modelo de lenguaje opera en un espacio distinto, generalmente uno de mayor dimensión, entonces se incorpora una capa de proyección, también referenciada como adaptador o conector. Esta capa corresponde a una red MLP que proyecta los embeddings visuales desde dimensión

512 al tamaño del embedding del LLM, que en el caso del modelo seleccionado es 896. Esta capa, al tener parámetros entrenables, también ajusta sus parámetros para que el LLM haga mejores predicciones. Este módulo permite una concatenación directa y coherente entre la información visual y textual.

La elección del LLM está fuertemente condicionada por dos factores principales: la longitud del contexto requerida y las restricciones de memoria GPU. En este trabajo, la secuencia de entrada al LLM está dominada por los embeddings de parches, pudiendo alcanzar decenas de miles de tokens. En consecuencia, el modelo seleccionado debe soportar longitudes de contexto del orden de al menos 10.000 tokens. Esta restricción reduce considerablemente el conjunto de modelos disponibles, excluyendo la mayoría de los LLMs convencionales con contextos entre 2.048 y 8.192 tokens. Entre los modelos públicamente disponibles con soporte para contextos largos y buen desempeño general son las familias LLaMA 3.2, Qwen 2.5 y MedGemma. Adicionalmente, existe un compromiso directo entre el tamaño del modelo de lenguaje y el número de parches de la WSI que pueden ser procesados simultáneamente. Modelos de mayor tamaño, como LLMs de 7B parámetros, obligan a reducir drásticamente el número de parches considerados (por ejemplo, a 2.000 parches), llegando a poder representar menos del 20 % de la información total de una WSI y omitir regiones histopatológicamente relevantes. En contraste, el uso de modelos de menor tamaño permite aumentar significativamente la cobertura espacial de la WSI. En este trabajo, se prioriza la inclusión de la mayor cantidad posible de parches por WSI, alcanzando una configuración estable de hasta 10.000 parches mediante el uso de un LLM de 0.5B parámetros. Con ello tanto en inferencia como en entrenamiento se realiza un muestreo uniforme de los parches para las WSI que contengan más de 10.000 parches. Bajo este criterio, se selecciona Qwen2.5-0.5B-Instruct, el cual no solo permite una mayor cobertura de la WSI, sino que además presenta un entrenamiento más estable bajo las restricciones de memoria disponibles. En la Figura 2.3 se puede ver de forma general cada módulo explicado del LLM multimodal.

3.3. Dataset

Dado que el objetivo del modelo es generar respuestas y reportes textuales condicionados a una WSI, se requiere un dataset de tipo *instruction tuning*, compuesto por

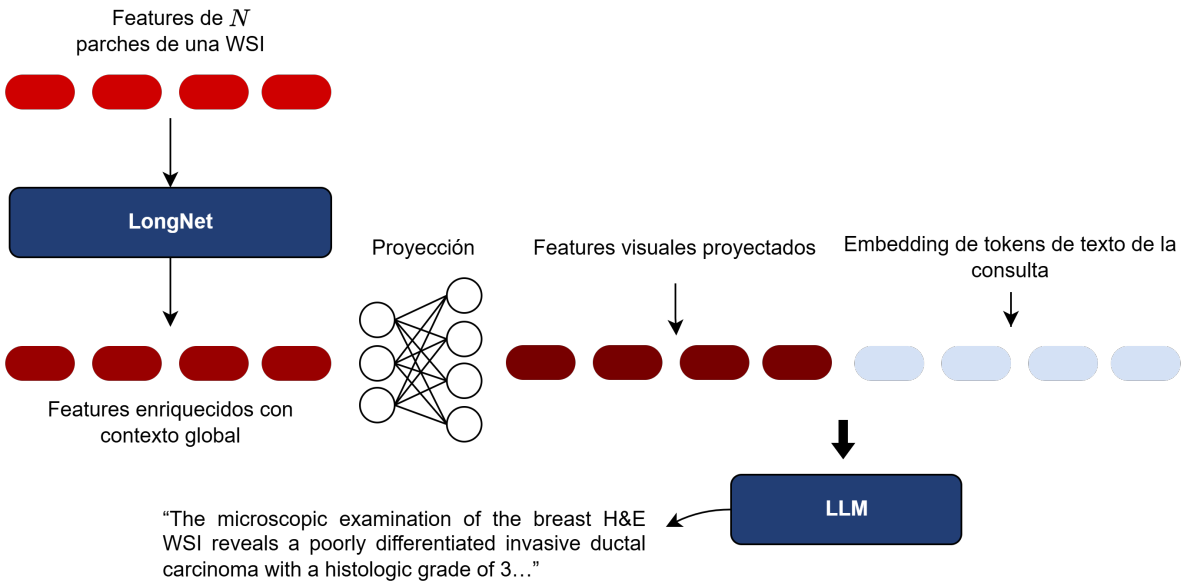


Figura 2.3: Pipeline completo del LLM multimodal

una instrucción en lenguaje natural y su correspondiente respuesta. Adicionalmente, la entrada debe incluir representaciones visuales de la WSI en forma de embeddings de parches, los cuales forman parte explícita del contexto de entrada del modelo.

Bajo estos requerimientos, cada muestra del dataset adopta la siguiente estructura:

```
{
  "id": "train::1",
  "image": [
    "./BRCA/TCGA-BH-A0BD-01Z-00-DX1.h5"
  ],
  "conversations": [
    {
      "from": "human",
      "value": "Generate a precise summary ....\n<image>"
    },
    {
      "from": "gpt",
      "value": "The pathological findings indicate a diagnosis..."
    }
```

```
    ]
  }
```

En este formato, el campo `image` referencia un archivo que contiene los embeddings de los parches de la WSI, mientras que el campo `conversations` define el par instrucción-respuesta utilizado para el entrenamiento supervisado del modelo.

Las WSIs utilizadas en este trabajo provienen del repositorio público The Cancer Genome Atlas (TCGA), el cual contiene miles de diapositivas histopatológicas teñidas con H&E, junto con metadatos clínicos asociados. En particular, se utilizan únicamente WSIs correspondientes a cáncer de mama (BRCA). Adicionalmente, se consideran datasets derivados del TCGA reportados en la literatura, específicamente SlideInstruct y WSI-Bench. Estos datasets proporcionan pares instrucción-respuesta asociados a WSIs, lo que los convierte en una base adecuada para tareas de *instruction tuning* multimodal. Si bien SlideInstruct y WSI-Bench constituyen una fuente valiosa de datos, ambos incluyen información que no es directamente inferible a partir de una WSI teñida con H&E. En particular, contienen respuestas relacionadas con *staging*, análisis molecular y otros atributos clínicos que no son observables exclusivamente a partir de la imagen histopatológica utilizada. El uso de este tipo de información introduciría ruido no deseable durante el entrenamiento.

Por esta razón, se aplican los siguientes criterios de filtrado:

- Se seleccionan únicamente muestras correspondientes a cáncer de mama.
- Se excluyen instrucciones y respuestas que requieran información no explícitamente observable en la WSI.
- Se eliminan muestras con respuestas excesivamente cortas, definiendo un umbral mínimo de 4 tokens.

Este último criterio responde a una decisión metodológica orientada a evitar que el modelo aprenda patrones de respuesta de una o pocas palabras, frecuentes en tareas de clasificación, y que no son representativos del objetivo de generación de reportes descriptivos y respuestas a preguntas abiertas.

Habiendo curado ambos datasets y fusionándolos apropiadamente, se crean dos datasets. Uno del estilo WSI-Reporte, orientado a que el modelo aprenda a redactar y procesar consultas relacionadas a la generación de reporte. El segundo siendo de VQA,

para que el modelo pueda responder consultas específicas de la WSI. En particular, el primer dataset contiene pares WSI-reporte, esto es una consulta en lenguaje natural del estilo: "Write a brief account that captures the main conclusions of the pathological evaluation from the whole slide image", y como respuesta el reporte de la WSI. Para el dataset de VQA se tiene pares de preguntas mas específicas como por ejemplo: "How many mitoses per 10 HPF are observed in the invasive ductal carcinoma?" y como respuesta: "2 mitoses per 10 high-power fields (HPF) are observed in the invasive ductal carcinoma.". Todo el dataset se compone de 1126 WSI únicas de TCGA en cáncer de mama.

En el Cuadro 3.1 se muestran la cantidad de muestras totales por tipo de dataset.

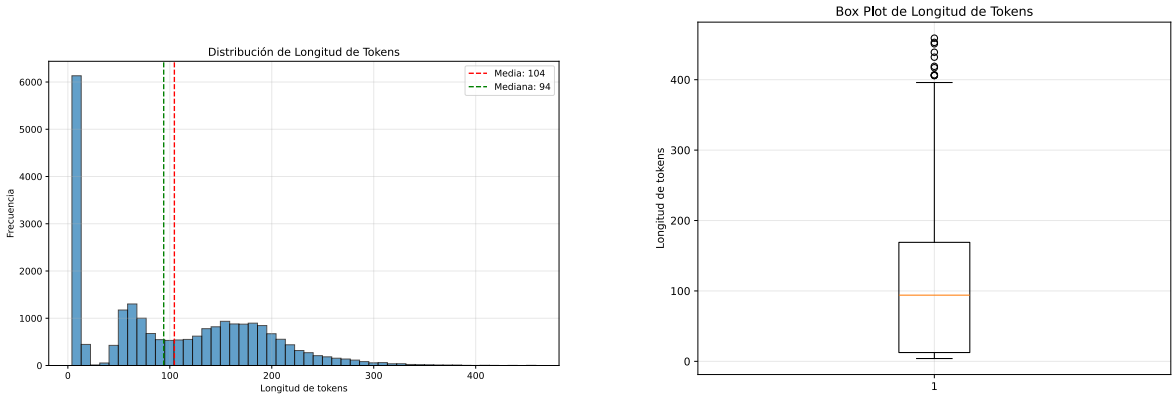
Cuadro 3.1: Resumen del número de muestras de los datasets utilizados.

Dataset	**Número de muestras**
WSI–Reporte	1989
VQA	21474

Notar que al fusionar y curar ambos datasets públicos, se observaron casos en los que una misma WSI tenía asociada dos instrucciones y reportes redactados de manera distinta. Esta característica, fue revisada manualmente y no se consideró perjudicial para el entrenamiento, ya que favorece la robustez semántica del modelo al exponerlo a variaciones lingüísticas del mismo contenido clínico. En particular hubieron 864 WSI duplicadas. No obstante, se aseguró que las particiones de entrenamiento, validación y prueba se realizaran a nivel de WSI para evitar fuga de información visual.

La Figura 3.4 muestra un histograma y un boxplot de la distribución de las respuestas del dataset en función de la cantidad de tokens. Se evidencia una distribución asimétrica positiva, con una media de 104.47 tokens y una mediana de 94 tokens, lo que indica la presencia de una cola derecha influenciada por respuestas extensas. La desviación estándar (81.75 tokens) confirma una variabilidad considerable en la extensión textual. El 50 % de las respuestas se concentra entre 12 y 169 tokens, mientras que el 90 % no supera los 211 tokens y el 99 % se mantiene bajo 304 tokens, lo que sugiere que los casos extremadamente largos son poco frecuentes. En términos de distribución por intervalos, la mayor proporción de muestras (35.55 %) se ubica entre 100 y 200 tokens, seguida por el rango 0–50 tokens (30.12 %), evidenciando la coexistencia de respuestas breves y descripciones más extensas correspondientes a reportes histopatológicos. No se observan respuestas superiores a 500 tokens, lo que indica que el dataset presenta un

límite natural de longitud compatible con configuraciones estándar de entrenamiento sin requerir truncamientos agresivos.



(a) Histograma de distribución de tokens. **(b)** Boxplot de longitud de respuestas.

Figura 3.4: Distribución de la longitud de las respuestas en el dataset.

3.4. Estrategia de Entrenamiento

El entrenamiento se divide en dos etapas: etapa 1 y etapa 2.

La primera etapa tiene como objetivo entrenar el *slide-level encoder* (LongNet) y la capa de proyección, manteniendo el LLM congelado. Siguiendo la estructura del dataset, se busca que tanto LongNet como la capa de proyección ajusten sus pesos para minimizar la pérdida de la función de *cross-entropy* en cada predicción del LLM, tomando como entrada las características de los parches y la instrucción o consulta asociada.

El dataset utilizado contempla pares WSI-reporte y/o pares WSI-*caption*, lo que permite al codificador contextual y a la capa de proyección capturar una representación global de la WSI. De esta forma, el modelo aprende a mapear información visual de alta dimensión hacia un espacio compatible con el modelo de lenguaje.

El entrenamiento se realiza hasta que la pérdida de validación deja de mejorar y comience a evidenciarse sobreajuste. Una vez se obtenga la mejor época de la etapa 1 se guarda el *checkpoint* y se continúa con la etapa 2 del entrenamiento.

En la segunda etapa se congela el codificador visual entrenado previamente y se descongelan los pesos del LLM, realizando ajuste fino mediante QLoRA. Esta técnica permite modificar únicamente un subconjunto de parámetros del modelo de lenguaje mediante

adaptadores de bajo rango, reduciendo el consumo de memoria y los requerimientos computacionales. El dataset utilizado es una combinación uniformemente espaciada de los dataset de WSI-Reporte y VQA, con el fin de que el modelo final se capaz de realizar las tareas de VQA y generación de reporte.

Se empleó el optimizador AdamW con regularización por *weight decay*. Se definieron tasas de aprendizaje diferenciadas para cada módulo. El proyector visual se entrenó con una tasa de aprendizaje de 5×10^{-5}. El LLM se entrenó con una tasa de aprendizaje de 1×10^{-4}. Se utilizó una fase de *warm-up* lineal durante la primera mitad de una época, comenzando en el 1 % de la tasa de aprendizaje y aumentando progresivamente hasta el valor máximo. Posteriormente la tasa se mantuvo constante. Para QLoRA se utilizó un rango $r = 8$ y un factor de escalamiento $\alpha = 16$, manteniendo la relación $\alpha = 2r$, configuración que ha demostrado buen desempeño en trabajos previos [41] [42].

Valores menores de r tienden a mejorar la generalización del modelo a costa de una menor capacidad representacional, mientras que valores mayores aumentan el riesgo de sobreajuste. En este trabajo se seleccionó $r = 8$ como compromiso entre capacidad y estabilidad.

Adicionalmente, se implementó *gradient checkpointing* en el modelo de lenguaje. Esta técnica permite recalcular activaciones intermedias durante el *backpropagation*, reduciendo el uso de memoria a costa de un mayor tiempo de cómputo. Se utilizó un tamaño de batch efectivo reducido de 1 muestra por iteración. Para compensar esta limitación se aplicó acumulación de gradientes durante 4 pasos consecutivos, logrando un batch efectivo de 4 muestras.

3.5. Configuración de la evaluación

La evaluación del sistema propuesto se diseña con el objetivo de medir tanto el desempeño linguistico como la de coherencia clínica de las respuestas generadas a nivel de *Whole Slide Images (WSI)*. Se consideran tareas de generación de reportes y consultas visuales (VQA). El proceso de evaluación combina métricas automáticas tradicionales del procesamiento de lenguaje natural con métodos de evaluación basados en LLM, adaptados al dominio medico.

El modelo se evalúa utilizando el *split* de *test* del dataset. Recordando que estos datasets

contienen pares de WSI-reporte y VQA. Además se evaluarán las dos etapas descritas previamente. En la primera etapa se evalúa el desempeño del encoder de contexto junto con la capa de proyección con el LLM congelado. En la segunda etapa se evalúa el primer *checkpoint* (las primeras 2862 muestras del dataset) y el modelo completo.

El desempeño cuantitativo se mide inicialmente mediante métricas estándar de generación de texto. Se utilizan BLEU para evaluar la coincidencia de n-gramas entre las predicciones y las referencias. Estas métricas permiten una comparación objetiva ya sea con trabajos similares de generación de reportes y entre los mismos *checkpoints* del entrenamiento.

Las métricas tradicionales presentan limitaciones al aplicarse al dominio médico. No capturan adecuadamente la validez medica de los conceptos mencionados. Penalizan variaciones semánticamente correctas pero textualmente distintas. No reflejan errores clínicamente críticos como confusiones diagnósticas o ausencias de hallazgos relevantes. Por esta razón, se complementan con métodos de evaluación más alineados al razonamiento clínico. Lo que se hace es incorporar un modelo LLM como evaluador automático de las respuestas generadas, algo que se conoce como *LLM-as-Judge*. Este enfoque se motiva por su capacidad de analizar coherencia semántica, pricisión clínica y relevancia contextual. El evaluador recibe como entrada la WSI descrita a nivel textual, la consulta o instrucción y la respuesta generada por el modelo. Se solicita al LLM emitir una puntuación y opcionalmente una breve justificación considerando criterios de corrección medica, completitud y claridad. La evaluación por LLM se estructura mediante una rúbrica que contempla exactitud diagnóstica, mención de hallazgos histopatológicos relevantes, coherencia global, completitud y nivel de alucinación. Cada criterio se puntúa en una escala de 1 a 10. Esta aproximación permite capturar matices clínicos que no son reflejados por las métricas clásicas de NLP. El LLM de razonamiento escogido para esta tarea es DeepSeek-R1-Distill-Qwen-32B, una versión reducida del modelo mas grande que consta de 671B de parámetros [43].

La rúbrica completa que recibe el LLM se puede ver en el Apéndice B

Los resultados cuantitativos se reportan como promedios sobre el conjunto de prueba para cada métrica tradicional y cada dimensión de la evaluación por LLM. Se presentan tablas comparativas entre las distintas etapas del entrenamiento. Se analizan tendencias generales de mejora o degradación del desempeño.

Se realiza una comparación directa entre los resultados obtenidos tras la primera etapa

y la segunda etapa de entrenamiento. Este análisis permite evaluar el impacto del fine-tuning del LLM mediante QLoRA sobre la calidad de generación. Se estudian variaciones tanto en métricas automáticas como en las puntuaciones clínicas otorgadas por el evaluador basado en LLM.

La evaluación cualitativa se incorpora con el objetivo de analizar la validez clínica y la coherencia semántica de las respuestas generadas por el modelo en el contexto de histopatología. Este análisis permite examinar aspectos que no son capturados por métricas automáticas, tales como la corrección conceptual de los hallazgos, la consistencia interna del reporte y la alineación con la información visual presente en la WSI. Se selecciona una WSI del conjunto de test y se le pedirá al modelo generar un reporte automático, ademas de hacerle preguntas acerca de la WSI. Asimismo, se analiza el impacto del fine-tuning del LLM en la calidad del lenguaje y en la precisión del contenido. Se comparan ejemplos generados tras la primera y segunda etapa de entrenamiento, destacando el cambio en fluidez, especificidad y coherencia clínica, así como posibles degradaciones asociadas a sobreajuste o generación excesivamente confiada. Finalmente, la evaluación cualitativa permite contextualizar los resultados cuantitativos y apoyar la interpretación de las métricas automáticas. Este análisis no pretende reemplazar la validación por expertos humanos, sino ofrecer una aproximación sistemática y reproducible para el estudio de modelos generativos aplicados a histopatología computacional.

Por último, cabe mencionar que la evaluación se centra en la concordancia, coherencia y fidelidad descriptiva entre las predicciones generadas por los modelos automáticos y los reportes de referencia, desde una perspectiva técnica y metodológica. Si bien se emplea terminología histopatológica estándar, este trabajo no tiene como objetivo reemplazar la interpretación diagnóstica realizada por especialistas en anatomía patológica. En consecuencia, pueden existir imprecisiones conceptuales menores que no afectan la validez del análisis comparativo, cuyo énfasis principal radica en la identificación de discrepancias estructurales, semánticas y clínicas relevantes entre predicción y ground truth.

CAPÍTULO 4

Resultados

En este capítulo se presenta el funcionamiento del modelo, los resultados cuantitativos, comparaciones relevantes, observaciones y análisis de los resultados.

4.1. Entrenamiento Etapa 1

Para la primera etapa, se hace entrenamiento de LongNet y la capa de proyección con el dataset de la etapa 1. Recordando que este incluía pares WSI-reporte. El LLM se encuentra congelado para este entrenamiento. Se entrena el modelo hasta que la pérdida del set de validación deje de mejorar. Esto se obtiene en la época número 9. En la Figura 1.1 se puede apreciar tanto el valor del error de la pérdida del conjunto de entrenamiento como el de validación.

Una de las formas mas comunes de interpretar este valor es a través de la métrica perplexity. Perplexity mide entre cuántos tokens candidatos probables está eligiendo el modelo, utilizando la siguiente formula:

$$Perplexity = 2^{H(P,Q)} \tag{1.1}$$

Recordando que $H(P, Q)$ es la cross entropy, que mide qué tan bien una distribución predicha Q se aproxima a la distribución verdadera P, el mejor valor de pérdida en

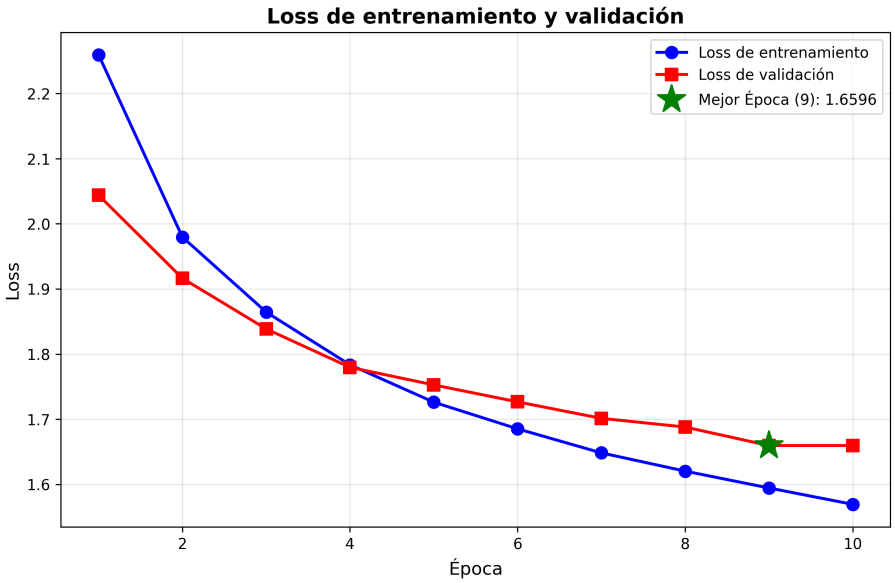


Figura 1.1: Valores de Loss para un entrenamiento de 10 épocas

validación fue de ≈ 1,66. Al reemplazar este valor en la ecuación correspondiente, se obtiene un perplexity de ≈ 3,16. Esto implica que, en promedio, el modelo mantiene una incertidumbre equivalente a elegir entre aproximadamente 3 y 4 tokens posibles en cada paso de generación. En consecuencia, el modelo presenta un nivel de incertidumbre relativamente bajo, lo que sugiere una buena capacidad de ajuste a la distribución del conjunto de validación y una generación lingüística razonablemente precisa y consistente.

4.2. Entrenamiento Etapa 2

Para la segunda etapa de entrenamiento, se mantiene LongNet congelado y se entrena tanto la capa de proyección como el LLM con QLoRA por una época con el dataset de la etapa 2 que contiene tanto VQA como pares WSI-reporte. Como se expuso en la sección anterior, esta etapa se evalúa con dos métricas principales: NLP tradicional (BLEU) y con un *LLM-as-a-judge*.

En los Cuadros 2.1 y 2.2 se pueden ver los resultados para las métricas BLEU-1 y BLEU-2.

Los resultados presentados en los Cuadros 2.1 y 2.2 muestran comportamientos diferen-

Cuadro 2.1: Comparación métrica BLEU entre distintas etapas de entrenamiento con el test de Etapa 1.

Etapa de entrenamiento	**BLEU-1**	**BLEU-2**
Etapa I: LLM-congelado	0.2626	0.1227
Etapa II: Proy. y LoRA LLM (1er checkpoint)	0.2299 (-12.45 %)	0.0928 (-24.40 %)
Etapa II: Proy. y LoRA LLM (Final)	0.2694 (+2.59 %)	0.1256 (+2.37 %)

Cuadro 2.2: Comparación métrica BLEU entre distintas etapas de entrenamiento con el test de Etapa 2 (solo VQA).

Etapa de entrenamiento	**BLEU-1**	**BLEU-2**
Etapa I: LLM-congelado	0.1443	0.0571
Etapa II: Proy. y LoRA LLM (1er checkpoint)	0.1493 (+3.41 %)	0.0655 (+14.69 %)
Etapa II: Proy. y LoRA LLM (Final)	0.1701 (+17.86 %)	0.0868 (+51.93 %)

ciados entre las etapas de entrenamiento dependiendo del conjunto de evaluación. En el test de la Etapa 1, el primer *checkpoint* de la segunda etapa evidencia una disminución en BLEU-1 y BLEU-2 respecto al modelo con LLM congelado, lo que sugiere una fase inicial de reajuste del espacio lingüístico producto del fine-tuning con LoRA. Sin embargo, el último *checkpoint* no solo recupera el desempeño previo, sino que lo supera levemente, indicando una adaptación progresiva y estable del modelo al nuevo esquema de entrenamiento conjunto. En contraste, al evaluar sobre el test de la Etapa 2 (solo VQA), se observa una mejora consistente desde el primer *checkpoint*, con incrementos más pronunciados en BLEU-2, lo que sugiere una mayor coherencia en secuencias de n-gramas más largas. El último *checkpoint* presenta una mejora sustancial, especialmente en BLEU-2 (+51.93 %), lo que indica una mejor alineación con la distribución textual del dataset multimodal introducido en la segunda etapa. En conjunto, estos resultados sugieren que el fine-tuning conjunto con QLoRA permite una adaptación efectiva al nuevo dominio, aunque con una transición inicial que puede afectar temporalmente el rendimiento en el dominio original. Estos resultados indican que el checkpoint final es el que mejor se desempeña para ambas tareas, VQA y generación de reporte.

No obstante, dado que BLEU evalúa coincidencia superficial de n-gramas y no necesariamente fidelidad clínica o coherencia semántica profunda, se complementa este análisis con una evaluación cuantitativa y cualitativa mediante un esquema *LLM-as-a-judge*. A través de un dataset privado se evaluaron 21 WSIs con sus correspondientes reportes.

En la Tabla 2.3 se pueden ver los promedios por cada categoría de la rúbrica del LLM evaluador (En el Apéndice B se pueden las definiciones de cada criterio).

Cuadro 2.3: Evaluación cualitativa mediante esquema LLM-as-a-Judge. Puntajes promedio por métrica y etapa de entrenamiento.

Métrica	**Configuración**	**Puntaje promedio**
Corrección factual	Etapa I: LLM-congelado	3.9048
	Etapa II: Proy. y LoRA LLM (1er checkpoint)	4.3333 (+10.98 %)
	Etapa II: Proy. y LoRA LLM (Final)	4.9048 (+25.61 %)
Relevancia	Etapa I: LLM-congelado	4.4286
	Etapa II: Proy. y LoRA LLM (1er checkpoint)	4.5714 (+3.23 %)
	Etapa II: Proy. y LoRA LLM (Final)	5.1429 (+16.13 %)
Completitud	Etapa I: LLM-congelado	3.2381
	Etapa II: Proy. y LoRA LLM (1er checkpoint)	3.7619 (+16.18 %)
	Etapa II: Proy. y LoRA LLM (Final)	4.7143 (+45.59 %)
Claridad	Etapa I: LLM-congelado	5.0000
	Etapa II: Proy. y LoRA LLM (1er checkpoint)	5.1905 (+3.81 %)
	Etapa II: Proy. y LoRA LLM (Final)	5.3333 (+6.67 %)
Nivel de alucinación	Etapa I: LLM-congelado	3.8571
	Etapa II: Proy. y LoRA LLM (1er checkpoint)	4.5238 (+17.28 %)
	Etapa II: Proy. y LoRA LLM (Final)	4.9524 (+28.40 %)

Los resultados obtenidos mediante el esquema *LLM-as-a-judge* muestran una mejora consistente y progresiva al avanzar desde la primera etapa con LLM congelado hacia el fine-tuning conjunto con QLoRA. Se observa un incremento sostenido en todas las métricas evaluadas, destacando particularmente la mejora en corrección factual (+25.61 %) y completitud (+45.59 %) en el checkpoint final, lo que sugiere una mejor alineación entre la representación visual y la generación textual clínicamente pertinente. Asimismo, el aumento en relevancia y claridad indica que el modelo no solo produce información más correcta, sino también mejor estructurada y más directamente relacionada con los hallazgos observados. En cuanto al nivel de alucinación, la mejora progresiva sugiere una reducción en la generación de contenido no sustentado, reforzando la estabilidad del modelo tras el entrenamiento conjunto. En conjunto, estos resultados complementan las métricas basadas en n-gramas y evidencian que el fine-tuning multimodal no solo mejora la similitud textual superficial, sino también la calidad semántica y clínica de los reportes generados.

Para el análisis cualitativo se seleccionó una de las muestras evaluadas previamente, con el objetivo de realizar una comparación directa entre la salida generada por el modelo y el reporte de referencia (ground truth). El modelo fue provisto con la consulta: "Generate a precise summary of the crucial findings derived from the pathology examination of the whole slide image." A partir de esta instrucción se obtuvo la siguiente predicción:

"The pathology report indicates that the whole slide image shows a ductal carcinoma in situ (DCIS) with central necrosis, cribriform patterns of invasion into surrounding stroma. There is evidence for lymphovascular spread to local structures and involvement by invasive cancer cells extending beyond margins. Additionally, there are areas within the DCIS showing focal micropapillary features consistent with microcalcifications or calcification due to i deposition on the tumor surface. These findings collectively suggest an aggressive breast lesion characterized as moderately differentiated according to Histologic Grade 2 criteria. The presence of vascular invasion further underscores this classification. Furthermore, there's no mention of regional lymph node dissection results from surgical biopsy sites but rather those adjacent to the initial excised specimen area. Overall assessment confirms these pathological characteristics, including grade II differentiation based upon histological analysis."

Siendo el ground truth: "Microscopic observation of the pathology slide reveals invasive epithelial neoplasia within breast tissue, characterized by infiltrating ducts and cords composed of cells exhibiting Grade 3 nuclear atypia, featuring significant variation in nuclear size and shape. Mitotic activity is elevated, with 13 mitoses counted in 10 high-power fields (HPFs), indicative of rapid cell proliferation. Adjacent to the invasive component, there is extensive carcinoma ductal in situ (DCIS) of high nuclear grade, displaying a cribriform pattern and associated comedonecrosis, representing approximately 5 % of the total tumor volume. Other benign changes noted include columnar metaplasia, usual ductal hyperplasia, cystic glandular dilations, and intraductal microcalcifications. Vascular invasion is present within the invasive carcinoma. The Elston-Ellis grading system was used, resulting in a histological grade II score (Nuclear grade: 3; Tubule formation: 2; Mitosis count: 2 points; Total score: 7), classifying the carcinoma as moderately differentiated based on its architectural features and mitotic rate."

Teniendo la predicción del modelo y el ground truth se puede comparar cualitativamente la predicción del modelo. Se dispone de la ayuda de GPT-5.2 para capturar las similitudes y diferencias que requieren conocimiento médico especializado.

La comparación entre ambos textos permite identificar coincidencias clínicamente relevantes. Tanto la predicción como el ground truth describen la presencia de carcinoma ductal in situ (DCIS), un patrón cribiforme y necrosis central (equivalente a comedonecrosis en este contexto). Asimismo, ambos reportes señalan invasión vascular y mencionan microcalcificaciones asociadas al componente intraductal. Finalmente, coinciden en la clasificación global como carcinoma moderadamente diferenciado (grado histológico II). No obstante, se observan discrepancias relevantes desde el punto de vista diagnóstico. La predicción omite información cuantitativa crítica, como el grado nuclear específico (grado 3) y el conteo mitótico (13 mitosis en 10 HPFs), elementos fundamentales en la gradación histológica. Además, introduce información no sustentada en el ground truth, como referencias a ganglios linfáticos y rasgos micropapilares focales. Por otro lado, no menciona hallazgos benignos asociados (metaplasia columnar, hiperplasia ductal usual, dilataciones quísticas), presentes en el reporte real.

Haciendo un análisis estructural, el reporte presenta una base inicial razonable para la identificación de hallazgos histopatológicos relevantes, integrando observaciones sobre arquitectura ductal, necrosis y posibles patrones de crecimiento tumoral. Como versión preliminar, el texto logra captar elementos clave que orientan hacia un diagnóstico. Se observan descripciones que combinan hallazgos in situ con referencias a invasión y extensión tumoral sin una delimitación explícita entre componentes, lo que sugiere la necesidad de mayor refinamiento conceptual en etapas posteriores. Asimismo, ciertos elementos contextuales, como la mención de márgenes quirúrgicos y ganglios linfáticos, aparecen de manera incipiente y no sistematizada, lo que es consistente con un reporte en fase de exploración. En conjunto, el texto cumple un rol orientativo adecuado como punto de partida, pero requiere una reorganización y precisiones adicionales para alcanzar mayor coherencia, trazabilidad y claridad exigidas en un reporte diagnóstico definitivo.

Para mas ejemplos de generación de reportes, consultar el Apéndice A.

CAPÍTULO 5

Conclusiones

En el presente trabajo se diseñó e implementó un modelo multimodal visión-lenguaje orientado a la generación de reportes histopatológicos a partir de Whole Slide Images (WSI), bajo restricciones computacionales explícitas y con énfasis en la coherencia clínica de las salidas generadas. La propuesta integra un encoder visual basado en CONCH, un agregador contextual de largo alcance mediante LongNet con Dilated Attention, y un modelo generativo Qwen2.5-0.5B-Instruct adaptado mediante QLoRA, permitiendo procesar hasta 10.000 parches por WSI dentro de los límites de memoria de una GPU de 24GB.

Desde el punto de vista arquitectónico, el trabajo demuestra que es posible capturar información local rica a nivel de parche y, simultáneamente, modelar dependencias globales sin incurrir en el costo cuadrático de la atención estándar. La incorporación de LongNet permitió mantener escalabilidad en secuencias de gran longitud, mientras que la estrategia de proyección y alineamiento dimensional facilitó la integración efectiva con el modelo de lenguaje. La decisión metodológica de priorizar cobertura espacial (mayor número de parches) por sobre el tamaño del LLM resultó adecuada en el contexto de análisis histopatológico, donde la omisión de regiones relevantes puede impactar directamente en la fidelidad clínica del reporte generado.

En términos de entrenamiento, la estrategia en dos etapas permitió desacoplar el aprendizaje de la representación visual contextual del ajuste lingüístico fino. La primera etapa alcanzó una pérdida de validación de aproximadamente 1.66, equivalente a una perplexity cercana a 3.16, indicando una modelación lingüística estable bajo el esquema con

LLM congelado. La segunda etapa, mediante fine-tuning con QLoRA, mostró una adaptación progresiva del modelo multimodal completo. Si bien se observó una disminución inicial en métricas BLEU sobre el conjunto original, el checkpoint final no solo recuperó el desempeño previo, sino que lo superó. En el conjunto de evaluación de la segunda etapa, se evidenció una mejora sustancial, particularmente en BLEU-2, lo que sugiere mayor coherencia en secuencias de mayor longitud.

Más relevante aún, la evaluación mediante el esquema LLM-as-a-Judge mostró mejoras consistentes en corrección factual, relevancia, completitud y reducción del nivel de alucinación. Destaca especialmente el incremento en completitud (+45.59 %) y corrección factual (+25.61 %) en el modelo final respecto a la configuración con LLM congelado. Estos resultados indican que el fine-tuning conjunto no solo mejora la similitud superficial con el texto de referencia, sino que fortalece la alineación semántica y clínica entre imagen y reporte generado.

El análisis cualitativo permitió identificar tanto fortalezas como limitaciones. El modelo logra capturar patrones histopatológicos centrales, como la presencia de DCIS, necrosis y gradación tumoral global, evidenciando una correcta transferencia de información visual hacia el espacio lingüístico. Sin embargo, aún persisten omisiones de detalles cuantitativos críticos e introducción ocasional de información no sustentada visualmente, lo que refleja desafíos inherentes a la generación automática en dominios clínicos de alta precisión.

En conjunto, los resultados validan la factibilidad técnica de entrenar un modelo multimodal de largo contexto para generación de reportes histopatológicos bajo recursos computacionales limitados, evidenciando mejoras medibles tanto en métricas automáticas como en evaluación clínica estructurada. No obstante, el sistema no reemplaza la interpretación diagnóstica experta, sino que se posiciona como una herramienta de apoyo potencial para tareas de resumen, estructuración preliminar de hallazgos o asistencia en revisión.

5.1. Trabajo futuro.

Una primera línea futura es la exploración de entrenamiento distribuido. Esto permitiría utilizar LLMs de mayor tamaño sin reducir la cantidad de parches procesados por WSI. Se podrían implementar esquemas como data parallelism, model parallelism. Esto permitiría estudiar el impacto real del tamaño del LLM en la coherencia clínica y el razonamiento diagnóstico.

Otra línea consiste en incorporar mecanismos de Retrieval-Augmented Generation (RAG). El modelo podría recibir contexto adicional además de los embeddings visuales. Este contexto podría provenir de información estructurada previamente extraída de la misma WSI. También podría incluir conocimiento externo como guías clínicas o criterios histopatológicos. Esto permitiría reducir alucinaciones y mejorar precisión factual.

Se propone además incorporar mapas de atención o mapas de calor sobre la WSI. Esto permitiría visualizar qué regiones influyen en cada respuesta generada. La interpretabilidad es un requisito clave en aplicaciones médicas. Esta extensión aumentaría transparencia y confianza en el sistema.

Finalmente, se plantea incorporar interactividad con especialistas humanos. El modelo podría operar bajo un esquema human-in-the-loop. Patólogos podrían corregir, refinar o cuestionar las respuestas generadas. Esta retroalimentación permitiría mejorar el modelo de forma continua. Además facilitaría su integración en entornos clínicos reales.

Bibliografía

[1] H. S. et al., "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] World Health Organization, "The top 10 causes of death," https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2023, accessed: 2026-02-21.

[3] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, 2015.

[4] Y. B. Y. LeCun and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[5] D. Li, G. Wan, X. Wu, X. Wu, X. Chen, Y. He, C. G. Lian, P. K. Sorger, Y. R. Semenov, and C. Zhao, "Multi-modal foundation models for computational pathology: A survey," 2025. [Online]. Available: https://arxiv.org/abs/2503.09091

[6] N. H. et al., "Breast cancer," *Nature Reviews Disease Primers*, vol. 5, no. 1, p. 66, 2019.

[7] National Cancer Institute, "The cancer genome atlas program," 2023, accessed: 2026-02-21. [Online]. Available: https://www.cancer.gov/tcga

[8] The Royal College of Pathologists, "Histopathology," 2023, accessed: 2026-01-12. [Online]. Available: https://www.rcpath.org/discover-pathology/news/fact-sheets/histopathology.html

[9] F. G. et al., "An analysis of pathologists' viewing processes as they diagnose whole slide digital images," *Journal of Pathology Informatics*, vol. 13, p. 100104, 2022.

[10] M. I. Samuelson, S. J. Chen, S. A. Boukhar, E. M. Schnieders, M. L. Walhof, A. M. Bellizzi, R. A. Robinson, and A. Rajan K D, "Rapid validation of whole-slide imaging for primary histopathology diagnosis: A roadmap for the sars-cov-2 pandemic era," *American Journal of Clinical Pathology*, vol. 155, no. 5, pp. 638–648, 01 2021. [Online]. Available: https://doi.org/10.1093/ajcp/aqaa280

[11] U.S. Food and Drug Administration, "Fda allows marketing of first whole slide imaging system for digital pathology," https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-whole-slide-imaging-system-digital-pathology, Apr 2017, accessed: 2026-02-21.

[12] E. A. et al., "Computational pathology definitions, best practices, and recommendations for regulatory guidance," *The Journal of Pathology*, vol. 249, no. 3, pp. 286–294, 2019.

[13] P. B. et al., "Qupath: Open source software for digital pathology image analysis," *Scientific Reports*, vol. 7, p. 16878, 2017.

[14] A. G. et al., "Openslide: A vendor-neutral software foundation for digital pathology," *Journal of Pathology Informatics*, vol. 4, p. 27, 2013.

[15] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[16] G. L. et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[18] A. D. et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[19] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber *et al.*, "A visual-language foundation model for computational pathology," *Nature Medicine*, vol. 30, p. 863–874, 2024.

[20] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," 2022. [Online]. Available: https://arxiv.org/abs/2205.01917

[21] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, B. Chen, A. Zhang, D. Shao, A. H. Song, M. Shaban *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, 2024.

[22] A. V. et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] A. Grattafiori, A. Dubey, A. Jauhri *et al.*, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[24] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2412.15115

[25] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej *et al.*, "Gemma 3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.19786

[26] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

[27] M. S. Seyfioglu, W. O. Ikezogwo, F. Ghezloo, R. Krishna, and L. Shapiro, "Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos," 2025. [Online]. Available: https://arxiv.org/abs/2312.04746

[28] Y. Liang, X. Lyu, W. Chen, M. Ding, J. Zhang, X. He, S. Wu, X. Xing, S. Yang, X. Wang, and L. Shen, "Wsi-llava: A multimodal large language model for whole slide image," 2025. [Online]. Available: https://arxiv.org/abs/2412.02141

[29] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly *et al.*, "Medgemma technical report," 2025. [Online]. Available: https://arxiv.org/abs/2507.05201

[30] Y. Chen, G. Wang, Y. Ji, Y. Li, J. Ye, T. Li, M. Hu, R. Yu, Y. Qiao, and J. He, "Slidechat: A large vision-language assistant for whole-slide pathology image understanding," 2025. [Online]. Available: https://arxiv.org/abs/2410.11761

[31] S. Afrin, M. Z. Haque, and A. Mastropaolo, "A systematic literature review of parameter-efficient fine-tuning for large code models," 2025. [Online]. Available: https://arxiv.org/abs/2504.21569

[32] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, and M. Tietz, "PEFT: State-of-the-art parameter-efficient fine-tuning methods," https://github.com/huggingface/peft, 2022.

[33] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *ACL*, 2021.

[34] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[35] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

[36] K. P. et al., "Bleu: a method for automatic evaluation of machine translation," *ACL*, 2002.

[37] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL Workshop*, 2004.

[38] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation," in *ACL Workshop*, 2005.

[39] H. Li *et al.*, "Llms-as-judges: A comprehensive survey on llm-based evaluation methods," https://arxiv.org/abs/2412.05579v2, 2024, arXiv preprint.

[40] J. Ding, S. Ma, L. Dong, X. Zhang, S. Huang, W. Wang, N. Zheng, and F. Wei, "Longnet: Scaling transformers to 1,000,000,000 tokens," 2023. [Online]. Available: https://arxiv.org/abs/2307.02486

[41] TabTune Documentation & DevTechTools, "Peft lora: Parameter-efficient fine-tuning for tabular models," 2024, heuristic recommendation: $\alpha = 2 \times r$ for balanced LoRA scaling. [Online]. Available: https://tabtune.lexsi.ai/advanced/peft-lora/

[42] Lightning AI, "Finetuning llms with lora and qlora: Insights from hundreds of experiments," https://lightning.ai/pages/community/lora-insights/, 2023, accessed: 2025-12-15.

[43] DeepSeek-AI, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.12948

APÉNDICE A

Ejemplos cualitativos adicionales de la generación de reporte.

A.1. Ejemplo 1

Predicción: The pathology report highlights the presence of invasive carcinoma with a grade 3 (G2) classification. This indicates poor differentiation, which is characterized by areas where cancer cells have lost their normal structure and function in forming glandular or tubule-like structures typical for healthy tissue architecture. The tumor exhibits an intermediate nuclear size-structure-potential score that suggests some variability but not extreme atypia compared to other grades like G1 or G0. Additionally, there are focal regions showing necrosis within ducts, as well as single foci exhibiting poorly differentiated features such as papillary growth patterns on MRI images without evidence of invasion into surrounding tissues further supporting this grading system's usefulness. Furthermore, no lymphovascular involvement was noted throughout these findings, indicating limited spread beyond local sites due primarily to infiltration rather than vascular dissemination.

Ground Truth: Infiltration of mammary glandular tissue by an invasive carcinoma characterized as very hypocellular, featuring scarce clusters and isolated atypical cells within the stroma; these atypical epithelial cells demonstrate positivity for pancytokeratin (AE1-AE3) and estrogen receptors (+++). The overall architectural pattern shows deviation from normal mammary gland structure due to the presence of infil-

trating neoplastic cells. Nuclear morphology exhibits atypia, although specific pleomorphism scores are not provided. Mitotic activity appears low, consistent with the v̈ery hypocellulard̈escription. Necrosis, calcification, and vascular invasion are not explicitly mentioned but the presence of invasive carcinoma suggests potential stromal involvement. This represents an invasive ductal carcinoma, likely Grade 1 based on the hypocellular nature and scarcity of atypical cells, though a formal grading score is not specified in the report.

La predicción del modelo y el ground truth coinciden en identificar la presencia de un carcinoma invasor mamario y en reconocer alteraciones de la arquitectura glandular normal asociadas a infiltración neoplásica, así como la existencia de atipia nuclear. No obstante, se observan discrepancias relevantes en la caracterización histopatológica. En particular, el modelo sobreestima el grado histológico al clasificar la lesión como pobremente diferenciada, mientras que el ground truth sugiere un tumor de bajo grado, con marcada hipocelularidad y baja actividad mitótica. Asimismo, la predicción introduce hallazgos no sustentados por el ground truth, tales como necrosis ductal, patrones papilares y referencias a imágenes por resonancia magnética, lo que evidencia confusión entre modalidades diagnósticas. Finalmente, el modelo omite características clave descritas en el ground truth, incluyendo la hipocelularidad tumoral y los resultados de inmunohistoquímica, limitando la fidelidad clínica de la predicción.

A.2. Ejemplo 2

Predicción: The pathology report indicates a diagnosis of invasive ductal carcinoma, classified as grade 2. This grading system evaluates the degree of differentiation and aggressiveness by assessing tubule formation (tubular/angioinvasion), nuclear pleomorphism, mitotic count, and lymphovascular invasion. In this case, there is moderate to high-grade intraductal components with areas showing necrosis in both solid and cribriform patterns within these tumors. The presence of extensive angiolymphatic invasion further underscores the aggressive nature of the tumor. Additionally, multiple foci exhibit multifocal extension into adjacent structures like skin and blood vessels without evidence for vascular or perineural involvement noted at any stage.

Ground Truth: Invasive ductal carcinoma characterized by high grade morphology. The tumor exhibits Grade 3 nuclear features, indicating significant pleomorphism and irre-

gularity in nuclear size and shape. Mitotic activity is elevated, with 18 mitoses counted per 10 high power fields, reflecting a high rate of cell division. Tubular formation is absent, indicated by a score of 0 %, signifying a lack of differentiated gland structures typical of normal breast tissue. These features contribute to an overall histological grade of 3, based on a Nottingham Grading System score of 8. Lymphovascular invasion is present within the lymph nodes examined. Metastatic disease is noted in four out of sixteen lymph nodes, one of which shows extranodal extension.

Si bien ambos textos coinciden en el diagnóstico de carcinoma ductal invasivo y en la presencia de invasión linfovascular, existe una discrepancia crítica en la gradación histológica: el ground truth reporta un grado 3 según el sistema de Nottingham (alto grado, 18 mitosis/10 HPF, ausencia de formación tubular), mientras que la predicción clasifica el tumor como grado 2.

APÉNDICE B

Rúbrica del LLM de razonamiento.

```
RUBRIC = """
You are evaluating answers from a small pathology AI model (0.5B parameters).

Compare the MODEL ANSWER to the GROUND TRUTH answer.
This model has limited capacity, so focus on:
- Does it capture the main point?
- Is it factually correct (even if less detailed)?
- Does it avoid major hallucinations?

Score each criterion on 0-10 scale:
- 0 = completely wrong
- 5 = partially correct, missing key details
- 10 = captures main point correctly

---

EVALUATION CRITERIA:

1. Factual Correctness (0-10)
Are the main facts correct?
(e.g., correct diagnosis, correct features mentioned)
Penalize factual errors heavily. Reward partial correctness.
```

```
2. Relevance (0-10)
Does the answer address the question asked?
10 = Directly answers the question
0 = Completely off-topic or generic

3. Completeness (0-10)
Does it include the most important information from GT?
This is a small model, so expect less detail.
Score based on KEY information only.

4. Clarity (0-10)
Is the answer coherent and understandable?
Penalize garbled or contradictory statements.

5. Hallucination Level (0-10)
10 = No major fabrications
5 = Minor unsupported details
0 = Major hallucinations (inventing diagnoses, features not in GT)

---

OUTPUT FORMAT:
Return scores in this exact format:

1. Factual Correctness: Score: X
2. Relevance: Score: X
3. Completeness: Score: X
4. Clarity: Score: X
5. Hallucination Level: Score: X

Replace X with scores 0-10. No other text needed.
"""
```