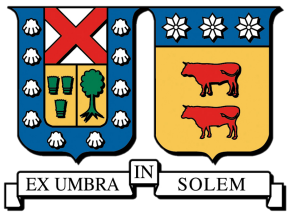


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA

Tesis de magíster

para obtener el título de
Magíster en Ciencias de la Ingeniería Informática

Evaluación de GPT como oráculo textual para clasificadores de tomografías computarizadas

Joaquín De Ferrari González

Composición del Jurado

Supervisado por:	Ph.D. Ricardo Ñanculef	Universidad Técnica Federico Santa María, Valparaíso Chile
Evaluador Interno:	Ph.D. Roberto Asín	Universidad Técnica Federico Santa María, Valparaíso Chile
Evaluador Externo:	Ph.D. Jocelyn Dunstan	Pontificia Universidad Católica de Chile, Santiago Chile

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: Evaluación de GPT como Oráculo Textual para Clasificadores de Tomografías Computarizadas

Nombre del candidato(a): Joaquín De Ferrari González

Carrera / Grado: Magíster en ciencias de la ingeniería informática

Campus: San Joaquín **Departamento:** Informática

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Ricardo Ñanculef Alegría ,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):

_____

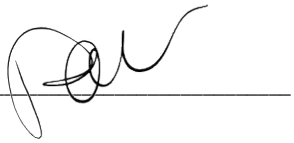
_____

_____

4.- FIRMAS

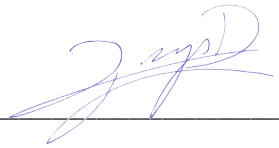
Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 18/05/05 **Firma:** _____



Estudiante o Candidato(a):

Fecha: 18/05/05 **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.

Declaración de autoría

Por la presente declaro que he redactado esta tesis sobre el tema

Evaluación de GPT como oráculo textual para clasificadores de tomografías computarizadas

de forma independiente. No he utilizado ninguna otra ayuda, fuente, figura o recurso que los declarados en las referencias bibliográficas. He identificado claramente todas las frases que han sido tomados de otras fuentes y los he citado correctamente.

Asimismo, declaro que —a mi leal saber y entender— este trabajo o partes del mismo no han sido presentados anteriormente, ni por mí ni por terceros, en esta ni en ninguna otra universidad.

Joaquín De Ferrari González

Valparaíso, 5 de mayo de 2026

Agradecimientos

El autor expresa su gratitud al Departamento de Informática de la Universidad Técnica Federico Santa María por el apoyo financiero brindado para la asistencia y presentación de un trabajo en la conferencia internacional *Artificial Intelligence in Medicine* (AIME 2025), celebrada en Pavía, Italia, del 23 al 26 de junio de 2025. Asimismo, se agradece el apoyo adicional para traslados gentilmente proporcionado por el profesor Mauricio Araya a través de AC3E ANID-Basal Project AFB240002.

Resumen

La escasez de tomografías computarizadas (TC) de tórax anotadas restringe el entrenamiento de sistemas diagnósticos. Esta investigación evalúa la supervisión débil mediante grandes modelos de lenguaje (LLM) para generar etiquetas clínicas a partir de informes radiológicos. El diseño desacopla la anotación del procesamiento visual. Un oráculo textual (`gpt-5-nano`) infiere etiquetas sobre un corpus masivo. Luego, clasificadores ligeros operan sobre características latentes extraídas por el codificador congelado de CT-CLIP, lo que elude el entrenamiento *end-to-end*. El estudio aplica *linear probing* e inferencia estadística emparejada con validación cruzada para mitigar sesgos de partición. El oráculo alcanzó alta fidelidad (F1-macro 0.8889) y estabilidad (κ = 0.947). En la clasificación visual, la supervisión manual lidera solo en regímenes reducidos (N = 50). Desde N = 100, la supervisión sintética supera estadísticamente a la manual en AUPRC (0.518 frente a 0.499) , y logra un techo asintótico de 0.568 en N = 10000. Al compensar el ruido del etiquetado con un mayor volumen de observaciones, la supervisión débil iguala y supera a la anotación experta bajo restricciones de hardware. No obstante, el estancamiento asintótico confirma un cuello de botella representacional inherente a la separabilidad lineal de las características estáticas de CT-CLIP.

Abstract

The scarcity of annotated chest computed tomography (CT) scans restricts the training of diagnostic systems. This research evaluates weak supervision using large language models (LLM) to generate clinical labels from radiology reports. The design decouples annotation from visual processing. A text oracle (`gpt-5-nano`) infers labels over a massive corpus. Then, lightweight classifiers operate on latent features extracted by the frozen CT-CLIP encoder, circumventing *end-to-end* training. The study applies *linear probing* and paired statistical inference with cross-validation to mitigate partition biases. The oracle achieved high fidelity (macro-F1 0.8889) and stability (κ = 0.947). In visual classification, manual supervision leads only in limited regimes (N = 50). From N = 100, synthetic supervision statistically outperforms the expert baseline in AUPRC (0.518 compared to 0.499) , reaching an asymptotic ceiling of 0.568 at N = 10000. By compensating for labeling noise with larger data volumes, weak supervision matches and surpasses expert annotation under hardware constraints. However, the asymptotic stagnation confirms a representational bottleneck inherent to the linear separability of static CT-CLIP features.

Índice

Índice de tablas

Índice de figuras

1

Introducción

1.1 Contexto clínico y tecnológico

El desarrollo de sistemas de aprendizaje profundo para el análisis clínico requiere grandes volúmenes de datos. En la práctica radiológica contemporánea, la tomografía computarizada (TC) constituye una herramienta indispensable para el diagnóstico morfológico. Sin embargo, existe una marcada disparidad en la disponibilidad de datos para el entrenamiento de modelos según la modalidad clínica.

Las modalidades volumétricas tridimensionales presentan una escasez de anotaciones estructuradas en comparación con las imágenes bidimensionales tradicionales. El proceso de etiquetado manual a nivel de vóxel o de volumen impone un alto costo cognitivo y temporal a los radiólogos expertos. Esta restricción logística y económica limita la cantidad de datos disponibles para el entrenamiento de arquitecturas supervisadas, lo que genera una barrera fundamental que impide el desarrollo de sistemas de diagnóstico automatizado robustos y escalables para volúmenes médicos complejos.

1.2 Planteamiento del problema

Frente a la escasez de datos etiquetados manualmente, el avance reciente en el procesamiento de lenguaje natural sugiere una vía metodológica alternativa. En los sistemas de información radiológica existe un volumen masivo de datos que empareja las tomografías con sus respectivos informes clínicos en texto libre. El uso de LLM permite operar como un oráculo de extracción capaz de interpretar la sintaxis y ambigüedad de estos informes radiológicos, y generar etiquetas clínicas con alta fidelidad en esquemas *zero-shot* y *few-shot*.

El desafío radica en transferir esta extracción de conocimiento del dominio textual al dominio visual

automatizado. Los modelos fundacionales visuales 3D ofrecen representaciones robustas mediante el alineamiento contrastivo lenguaje-imagen, pero el ajuste fino (*fine-tuning*) completo de estos codificadores exige altos recursos computacionales. La alta dimensionalidad del tensor espacial satura la memoria de los aceleradores de hardware durante la retropropagación, lo que vuelve este enfoque prohibitivo para la mayoría de los entornos clínicos e investigativos.

Para eludir esta restricción, la presente investigación plantea una estrategia que desacopla la generación de etiquetas del entrenamiento visual: utilizar características congeladas (*frozen features*) pre-computadas $\mathbf{z}$ y emplear un esquema de supervisión débil en el que un LLM genera etiquetas para entrenar un clasificador que puede ser desplegado con pocos recursos computacionales. En este esquema, un modelo de lenguaje actúa como maestro anotador sobre un gran volumen de datos no etiquetados, y un clasificador ligero actúa como estudiante operando sobre el espacio latente visual. En consecuencia, el problema científico central consiste en determinar si la supervisión débil generada a gran escala por modelos paramétricamente eficientes como `gpt-5-nano` permite entrenar clasificadores cuyo rendimiento iguale o supere al de modelos equivalentes entrenados con datos anotados por expertos.

1.3 Objetivos de la investigación

Objetivo general

Comparar el rendimiento de clasificadores visuales multi-etiqueta entrenados con imágenes TC de tórax anotadas automáticamente vía informe por un LLM frente a etiquetas obtenidas manualmente por expertos, operando sobre representaciones volumétricas pre-computadas.

Objetivos específicos

1. **Evaluar** la calidad, estabilidad y consistencia de las etiquetas generadas por el LLM respecto a las anotaciones manuales.
2. **Determinar** si el rendimiento está limitado por la separabilidad del espacio latente, comparando linear probing con MLPs bajo distintos umbrales.
3. **Caracterizar** las curvas de escalamiento de ambas fuentes de supervisión, identificando el techo asintótico y el punto de cruce entre supervisión débil y manual.

1.4 Hipótesis de trabajo

La formulación del problema y el diseño metodológico propuesto se sustentan en las siguientes hipótesis de trabajo:

- **Hipótesis 1 (Cuello de botella representacional):** En la clasificación de patologías torácicas sobre TC 3D, las representaciones pre-computadas de CT-CLIP imponen un techo de rendimiento que no se supera aumentando el volumen de datos ni la capacidad del clasificador.
- **Hipótesis 2 (Dinámica de escalamiento y supervisión débil):** Dentro del límite impuesto por las representaciones pre-computadas, existe un punto de cruce a partir del cual la supervisión débil mediante LLM iguala y supera a la supervisión experta, compensando el ruido de las etiquetas automáticas con un mayor volumen de datos.

2

Marco teórico

2.1 Dominio clínico y representación volumétrica

2.1.1 Tomografía computarizada

El análisis automatizado del tórax requiere comprender la estructura de las modalidades radiológicas volumétricas. Como describe Hounsfield (1973), la tomografía computarizada (TC) genera representaciones tridimensionales de la anatomía interna mediante la medición de la atenuación de rayos X. Desde una perspectiva computacional, un estudio de TC se define como un tensor tridimensional $\mathbf{X} \in \mathbb{R}^{D \times H \times W}$ compuesto por vóxeles.

Según explica Buzug (2008), cada vóxel representa la radiodensidad del tejido expresada en unidades Hounsfield (HU). Litjens et al. (2017) advierten que esta dimensionalidad espacial contiene la información morfológica necesaria para identificar patologías como nódulos, opacidades o calcificaciones, pero impone un costo de procesamiento significativamente mayor frente a las imágenes bidimensionales tradicionales.

2.1.2 Caracterización clínica de las patologías objetivo

La identificación de hallazgos en tomografía computarizada requiere una estandarización que asegure la consistencia diagnóstica. De acuerdo con el glosario de términos para imagen torácica de la *Fleischner Society* Hansell et al. (2008), las patologías seleccionadas se definen de la siguiente manera:

- **Calcificación de la pared arterial**: Presencia de depósitos de calcio en las túnicas de los vasos arteriales, manifestados como estructuras de alta atenuación en la tomografía computarizada. En

el contexto torácico, afecta predominantemente a la aorta y las arterias coronarias.

- **Linfadenomegalia**: Aumento del tamaño de uno o más ganglios linfáticos. Por convención radiológica, se considera patológico un diámetro del eje corto superior a 10 mm en las estaciones ganglionares mediastínicas e hiliares.
- **Nódulo pulmonar**: Opacidad redondeada o irregular, con márgenes bien o mal definidos, cuyo diámetro máximo es inferior o igual a 3 cm. Las lesiones que exceden este umbral se clasifican morfológicamente como masas.
- **Opacidad pulmonar**: Término genérico para cualquier área que atenúa el haz de fotones en la imagen radiológica, lo que reduce la transparencia del parénquima pulmonar. Este concepto engloba patrones específicos como la consolidación y la atenuación en vidrio esmerilado.
- **Secuela fibrótica pulmonar**: Manifestación de fibrosis pulmonar irreversible caracterizada por la presencia de reticulación, bronquiectasias por tracción y distorsión de la arquitectura pulmonar. Representa el estadio final de diversas agresiones al intersticio pulmonar.

2.2 Arquitecturas de aprendizaje profundo

Esta sección introduce las arquitecturas para procesamiento multimodal que permiten aprender representaciones vectoriales comunes texto-imagen.

2.2.1 Perceptrones multicapa (MLP)

Para abordar la clasificación final de características extraídas, Goodfellow et al. (2016) definen el MLP como la arquitectura fundamental de las redes neuronales artificiales. Opera mediante capas apiladas de nodos interconectados con funciones de activación no lineales. En el contexto de este trabajo, un MLP actúa como una función aproximadora $f_\theta : \mathbb{R}^d \to \{0,1\}^C$ que proyecta un vector latente estático $\mathbf{z} \in \mathbb{R}^d$ hacia un espacio de predicción multietiqueta de C clases. Como señalan Hastie et al. (2009), su bajo costo computacional permite entrenar y evaluar clasificadores directamente sobre representaciones precomputadas. Si bien son eficientes para la clasificación final sobre representaciones densas, los MLP resultan inadecuados para extraer atributos directamente desde datos crudos de alta dimensionalidad espacial o secuencial.

2.2.2 Transformers

Frente a las limitaciones de los perceptrones para modelar datos secuenciales y espaciales crudos, Vaswani et al. (2017) proponen la arquitectura *transformer*, fundamentada en el mecanismo de auto-atención (*self-attention*), la cual permite modelar dependencias a largo plazo superando las restricciones de arquitecturas previas. La formulación canónica de este mecanismo calcula la atención mediante la proyección lineal de las entradas hacia matrices de consultas ($\mathbf{Q}$), claves ($\mathbf{K}$) y valores ($\mathbf{V}$):

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

donde d_k representa la dimensión de las claves, el cual actúa como un factor de escalamiento en el denominador que estabiliza los gradientes de la función *softmax*. Este mecanismo pondera la relevancia de diferentes partes de la entrada de forma paralela. Como exponen Dosovitskiy et al. (2021); Hamamci et al. (2026), en el procesamiento de tomografías computarizadas (3D), la auto-atención se adapta estructuralmente —por ejemplo, mediante la factorización en ejes espaciales y de profundidad— para analizar eficientemente volúmenes médicos y modelar relaciones anatómicas globales.

2.2.3 Grandes modelos de lenguaje (LLM)

La capacidad de modelado de dependencias a largo plazo de los *transformers* posibilitó el desarrollo de los LLM en el dominio del NLP. Como detallan Brown et al. (2020); Touvron et al. (2023), modelos como los de la familia GPT se pre-entrenan sobre corpus masivos de texto y adquieren capacidades de razonamiento semántico y generalización algorítmica. En el entorno clínico, De Ferrari et al. (2025) demuestran que un LLM funciona como un oráculo de extracción $\Lambda_{\text{LLM}} : \mathcal{R} \to \{0,1\}^C$ capaz de interpretar la sintaxis y la ambigüedad de los informes radiológicos $r \in \mathcal{R}$. Esta traducción de texto libre a vectores de etiquetas estructuradas $\hat{\mathbf{y}}$ se logra sin actualizar los parámetros del modelo; la inferencia se condiciona a través de aprendizaje en contexto (*in-context learning*) bajo esquemas *few-shot*.

2.3 Alineación multimodal y modelos fundacionales

Los modelos fundacionales son arquitecturas de aprendizaje profundo pre-entrenadas a gran escala sobre datos diversos y no estructurados, diseñadas con una capacidad de generalización intrínseca que les permite adaptarse a una multiplicidad de tareas *downstream* sin requerir diseños arquitectónicos específicos para cada una. En el dominio de la visión y el lenguaje, el esquema de pre-entrenamiento contrastivo lenguaje-imagen (CLIP) constituye un modelo fundacional paradigmático. Radford et al. (2021)

establecen una correspondencia entre representaciones visuales y textuales mediante la proyección de ambas modalidades hacia un espacio latente compartido de dimensión d. Esta metodología facilita la adquisición de conceptos semánticos directamente desde supervisión en lenguaje natural y prescinde de etiquetas discretas durante el pre-entrenamiento.

2.3.1 El objetivo contrastivo (InfoNCE)

La alineación multimodal se operacionaliza mediante la minimización de una función de pérdida de entropía cruzada simétrica denominada InfoNCE (*Information Noise-Contrastive Estimation*) van den Oord et al. (2018). Dado un lote de B pares volumen-informe $(\mathbf{X}_i, r_i)$, el modelo emplea dos redes neuronales paramétricas independientes: un codificador visual que extrae *embeddings* volumétricos $\mathbf{z}_{v,i} = \Phi(\mathbf{X}_i)$, y un codificador de texto que extrae *embeddings* narrativos $\mathbf{z}_{t,i} = \Psi(r_i)$. El objetivo maximiza la similitud del coseno entre pares correspondientes (positivos) y la penaliza para los $B^2 - B$ pares no correspondientes (negativos) dentro del lote:

$$\mathcal{L} = \frac{1}{2B} \sum_{i=1}^{B} \left(-\log \frac{\exp(\text{sim}(\mathbf{z}_{v,i}, \mathbf{z}_{t,i})/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(\mathbf{z}_{v,i}, \mathbf{z}_{t,j})/\tau)} - \log \frac{\exp(\text{sim}(\mathbf{z}_{t,i}, \mathbf{z}_{v,i})/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(\mathbf{z}_{t,i}, \mathbf{z}_{v,j})/\tau)} \right) \tag{2.1}$$

donde $\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$ es la similitud del coseno y $\tau > 0$ es un parámetro de temperatura aprendible que ajusta la agudeza de la distribución de similitudes.

2.3.2 Especificidades de CT-CLIP

En la tomografía computarizada (TC), Hamamci et al. (2026) proponen el modelo CT-CLIP, el cual extiende este enfoque mediante el uso de *transformers* 3D factorizados. La factorización de la atención en componentes espaciales y de profundidad permite procesar tensores volumétricos $\mathbf{X} \in \mathbb{R}^{D \times H \times W}$ con mayor eficiencia computacional que las arquitecturas convolucionales tridimensionales estándar. Críticamente, a diferencia de los enfoques tradicionales que procesan reconstrucciones transversales independientes (*slice-by-slice*), esta arquitectura alinea la imagen volumétrica completa con el informe radiológico correspondiente.

La calidad de esta alineación reside en la capacidad del codificador visual para capturar información clínica relevante descrita en los informes, incluyendo su distribución anatómica global. Las representaciones resultantes —congeladas tras el pre-entrenamiento— son densas y linealmente separables. Esta característica permite el entrenamiento de clasificadores *downstream* para discriminar patologías sin requerir un *fine-tuning* importante de la arquitectura base, lo cual constituye el fundamento metodológico de esta investigación.

2.4 Mecanismos de adaptación de modelos fundacionales

Los modelos fundacionales requieren métodos de adaptación para aplicarlos a dominios específicos o entornos con recursos limitados.

2.4.1 *Fine-tuning*

El *fine-tuning* constituye el mecanismo tradicional para adaptar exhaustivamente modelos pre-entrenados a tareas específicas de un dominio clínico. El proceso requiere la actualización iterativa de la totalidad o una gran porción de los pesos de la red neuronal mediante la propagación de gradientes sobre un conjunto *gold standard*. Como explican Hamamci et al. (2026), en el procesamiento de volúmenes médicos tridimensionales, el *full fine-tuning* de arquitecturas fundacionales exige altos recursos computacionales. La alta dimensionalidad de los volúmenes 3D $\mathbf{X} \in \mathbb{R}^{D \times H \times W}$ satura la memoria de video (VRAM) de los aceleradores de hardware durante la fase de entrenamiento, lo que hace prohibitivo el *fine-tuning* completo en la mayoría de los entornos clínicos. El congelamiento de la arquitectura base elude esta barrera computacional al forzar a la red a operar únicamente como un extractor de vectores latentes estáticos $\mathbf{z}$, sobre los cuales se entrenan clasificadores de bajo costo sin pasar gradientes a las arquitecturas fundacionales.

2.4.2 *Linear probing*

El *linear probing* evalúa la calidad de las representaciones de un modelo congelado mediante una única capa lineal. A diferencia del *full fine-tuning*, este mecanismo restringe el aprendizaje a una única capa clasificadora puramente lineal (sin activaciones ocultas) acoplada sobre las representaciones latentes extraídas $\mathbf{z} \in \mathbb{R}^d$. Esta restricción aísla la calidad del espacio de features e impide que capas no lineales compensen representaciones poco separables.

En el análisis de generalización para modelos pre-entrenados, Radford et al. (2021) sugieren que la evaluación sistemática mediante *linear probing* permite cuantificar con solidez el grado real de separabilidad lineal de las clases objetivo en el espacio latente originalmente proyectado. Esto permite determinar si las limitaciones de rendimiento provienen del espacio de representaciones o de la capacidad del clasificador.

2.4.3 Ingeniería de prompts (*prompt engineering*)

El rendimiento de los LLM en tareas de extracción clínica depende del diseño del prompt. El *prompt engineering* permite alterar la salida del modelo hacia la tarea objetivo sin modificar sus pesos, una forma

nueva y económica de especialización a una tarea. En este ámbito, Brown et al. (2020) establecen que el aprendizaje *few-shot* condiciona la inferencia del oráculo textual Λ_{LLM} mediante la inclusión de pares de entrada y salida representativos dentro del propio contexto de consulta. Los ejemplos incluidos guían al modelo hacia el dominio radiológico y reducen las alucinaciones. Asimismo, la inclusión de restricciones negativas en el prompt reduce la ambigüedad entre patologías con terminología solapada. Estas técnicas de prompting permiten que modelos pequeños como `gpt-5-nano` alcancen rendimientos cercanos a los de modelos mayores a una fracción del costo.

2.5 Destilación de conocimiento y supervisión débil

2.5.1 Destilación de conocimiento

Como proponen Hinton et al. (2015), la destilación de conocimiento es un paradigma de aprendizaje automático diseñado para transferir la capacidad de generalización desde un modelo de alta complejidad (maestro) hacia un modelo más compacto y eficiente (estudiante). La formulación original minimiza la divergencia de Kullback-Leibler entre las distribuciones de probabilidad de salida de ambas redes. En escenarios de escasez de anotaciones verificadas, el paradigma se adapta hacia la supervisión débil multimodal donde un modelo de lenguaje masivo asume el rol de maestro y genera etiquetas sintéticas $\hat{\mathbf{y}}_i$ a partir de texto no estructurado. El modelo estudiante, típicamente un clasificador ligero como un perceptrón multicapa, aprende a mapear representaciones continuas $\mathbf{z}$ hacia el espacio de etiquetas discretas mediante la señal producida por el maestro. El clasificador visual aprende a predecir etiquetas sin requerir al LLM ni anotadores humanos durante la inferencia.

2.5.2 Supervisión débil

El entrenamiento de redes neuronales profundas requiere habitualmente grandes volúmenes de datos con anotaciones estructuradas revisadas por expertos. Para reducir esta dependencia, Zhou (2018) define la supervisión débil como un paradigma de aprendizaje automático que emplea fuentes de información ruidosas, heurísticas o modelos externos para generar etiquetas sintéticas $\hat{\mathbf{y}}_i$ de forma automatizada sobre grandes conjuntos de datos no etiquetados $\mathcal{D}_u = \{(\mathbf{X}_i, r_i)\}_{i=1}^{N}$.

Ratner et al. (2017) argumentan que en la intersección del análisis de imágenes médicas y el NLP, la narrativa clínica almacenada en los sistemas de información radiológica actúa como la fuente central de supervisión. La asignación de etiquetas generadas algorítmicamente permite entrenar modelos predictivos a gran escala. Kaplan et al. (2020) demuestran que la efectividad de este paradigma se evalúa

empíricamente mediante *scaling laws* para determinar si el volumen masivo de datos con etiquetas sintéticas logra compensar el ruido inherente y alcanzar el rendimiento de modelos entrenados con datos escasos de alta fidelidad.

2.5.3 Ensambles algorítmicos y mecanismos de consenso

La generación de etiquetas sintéticas mediante modelos de lenguaje introduce una varianza inherente debido a la naturaleza estocástica de la decodificación generativa. Los métodos de ensamble mitigan esta inestabilidad predictiva al agregar múltiples inferencias independientes sobre un mismo conjunto de datos no etiquetado $\mathcal{D}_u$. En el contexto de la supervisión débil multietiqueta, las predicciones de múltiples ejecuciones se agregan mediante reglas de consenso:

- **Regla de cualquier positivo (*any positive*)**: Asigna la clase si al menos una iteración del oráculo la detecta. Prioriza la sensibilidad y la recuperación de hallazgos ambiguos, a expensas de incrementar la tasa de falsos positivos en la señal de supervisión.
- **Voto mayoritario (*majority vote*)**: Asigna la clase final si estrictamente más de la mitad de las predicciones independientes concuerdan. Proporciona un equilibrio empírico entre sensibilidad y precisión, lo que suaviza anomalías aisladas.
- **Unanimidad (*unanimity*)**: Exige un acuerdo total entre todas las inferencias para asignar una etiqueta positiva. Maximiza la precisión de la etiqueta sintética y asegura una alta fidelidad semántica, pero penaliza la sensibilidad al descartar cualquier predicción con discordancia marginal.

Estos mecanismos estabilizan la señal de supervisión y permiten medir la fiabilidad del LLM antes de entrenar el clasificador visual.

2.5.4 Dinámica de las leyes de escalamiento

La evaluación empírica de modelos predictivos bajo distintos regímenes de volumen de datos obedece a leyes de escalamiento. Estas leyes modelan matemáticamente la relación entre la cantidad de ejemplos de entrenamiento y la capacidad de generalización del clasificador. En el contexto de la supervisión débil multimodal, donde se contrastan etiquetas sintéticas masivas contra anotaciones expertas escasas, esta dinámica se caracteriza por dos fenómenos críticos.

El techo asintótico (*asymptotic ceiling*) define el límite superior de rendimiento predictivo que un modelo alcanza al incrementar el volumen de datos de entrenamiento. Una vez superado cierto umbral de observaciones, la adición de nuevos ejemplos deja de aportar ganancias marginales significativas en la

reducción del error empírico. En arquitecturas de *frozen features*, este estancamiento refleja el límite de separabilidad lineal del espacio latente $\mathbf{z}$ y demuestra que el cuello de botella radica en la representación visual de entrada y no en el déficit de muestras de entrenamiento.

El *crossover point* cuantifica la cantidad exacta de datos requeridos para que un modelo entrenado con etiquetas débiles y ruidosas iguale o supere el rendimiento de un modelo equivalente entrenado con el *gold standard*. Este indicador permite evaluar si el volumen de datos débilmente supervisados compensa el ruido de las etiquetas automáticas.

2.6 Taxonomía de cuellos de botella en aprendizaje automático

La identificación de los factores que limitan la capacidad predictiva es un requisito fundamental para evaluar el rendimiento asintótico en modelos de aprendizaje automático. En arquitecturas donde la extracción de características es independiente del ajuste del clasificador final, estas restricciones se categorizan mediante una taxonomía de cuellos de botella:

- **Cuello de botella representacional**: Ocurre cuando las proyecciones latentes estáticas $\mathbf{z} \in \mathbb{R}^d$ carecen de la separabilidad lineal necesaria para discriminar las clases objetivo Alain and Bengio (2016). Esta limitante impone un techo al rendimiento e impide mejoras empíricas independientemente del volumen de datos o de la capacidad del clasificador.
- **Cuello de botella de etiquetas**: Surge por la introducción de ruido semántico o sesgos de calibración en la señal de entrenamiento Frénay and Verleysen (2013). En entornos de supervisión débil, la divergencia probabilística entre las etiquetas sintéticas $\hat{\mathbf{y}}_i$ y la distribución real (*gold standard*) restringe la generalización del modelo sobre datos de prueba independientes.
- **Cuello de botella de capacidad**: Se manifiesta cuando la complejidad del clasificador es insuficiente (*underfitting*) o excesiva (*overfitting*) respecto a la naturaleza de la tarea Goodfellow et al. (2016).
- **Cuello de botella de distribución**: Resulta de discrepancias estadísticas entre las particiones de entrenamiento y prueba, o de variaciones en la prevalencia poblacional de las patologías objetivo Moreno-Torres et al. (2012).
- **Cuello de botella de optimización**: Sucede cuando las dinámicas de convergencia o la selección de umbrales operativos impiden alcanzar el punto de decisión óptimo. Este escenario requiere ajustes posteriores a la inferencia probabilística para equilibrar la sensibilidad frente a la precisión en contextos de desbalance multiclase Lipton et al. (2014).

2.6.1 Optimización de umbrales de decisión (*threshold tuning*)

En la clasificación multi-etiqueta con desbalance de clases, es necesario desacoplar la capacidad discriminativa del modelo de la selección de umbrales de decisión Lipton et al. (2014). La inferencia de una red neuronal genera un vector de probabilidades continuas para cada clase objetivo, donde la asignación de la etiqueta binaria $\hat{\mathbf{y}} \in \{0,1\}^C$ depende de un umbral que, por defecto, se fija en $0{,}5$. *Threshold tuning* consiste en ajustar este valor de corte de forma independiente para cada patología sobre un conjunto de validación, con el fin de maximizar métricas sensibles al desbalance como el F1-score.

Este procedimiento permite aislar y mitigar el cuello de botella de optimización. De este modo, se garantiza que una representación latente con alta separabilidad —medida por métricas independientes del umbral como AUROC o AUPRC— no vea penalizado su rendimiento empírico por una partición subóptima del espacio de predicción.

3

Trabajo relacionado

Como señalan Diaz-Pinto et al. (2024), el desarrollo de sistemas de aprendizaje profundo en radiología requiere grandes volúmenes de datos anotados y la generación manual de estas etiquetas exige un alto costo temporal y cognitivo. La literatura explora diversas estrategias para extraer señales de supervisión de forma automatizada, las cuales han evolucionado desde sistemas basados en reglas hasta el uso de LLM como anotadores automáticos. Este trabajo combina supervisión débil con representaciones visuales 3D pre-computadas para operar bajo ambientes con baja capacidad computacional.

3.1 Evolución de la supervisión débil en análisis de imágenes médicas

Ante la escasez de imágenes médicas anotadas manualmente, investigaciones como las de Chng et al. (2023) y Dunnmon et al. (2020) destacan el uso de supervisión débil para extraer información estructurada directamente desde los informes radiológicos de texto libre. Los primeros enfoques estandarizados utilizaron sistemas basados en reglas lógicas y NLP. Irvin et al. (2019) ejemplifican esta etapa con CheXpert, el cual mapea menciones léxicas para radiografías de tórax e incorpora un manejo explícito de la incertidumbre. Como advierten Chng et al. (2023), la variabilidad sintáctica de la narrativa clínica evidenció la rigidez de las reglas manuales y motivó la transición hacia arquitecturas contextuales basadas en *transformers*. Investigaciones como las de Smit et al. (2020) con CheXbert y Yan et al. (2022) con RadBERT demostraron que modelos pre-entrenados ajustados sobre etiquetas generadas por reglas superan el rendimiento del entrenamiento supervisado tradicional ejecutado sobre conjuntos pequeños de datos curados por expertos.

La transferencia de estos enfoques bidimensionales a modalidades volumétricas tridimensionales, como la TC, requiere adaptaciones. Draelos et al. (2021) proponen herramientas específicas para este dominio, como SARLE, el cual predice múltiples anomalías simultáneas en TC mediante sistemas híbridos. Estas

variantes conservan una dependencia del diseño de reglas manuales para capturar patologías o hallazgos infrecuentes. Para superar esta limitación, enfoques posteriores propuestos por Dunnmon et al. (2020), como la programación de datos cruzados, integran funciones heurísticas con modelado generativo para automatizar la extracción.

3.2 LLM como oráculos clínicos

A nivel general, Brown et al. (2020) establecen que los LLM aplican fuertes capacidades de generalización en esquemas *zero-shot* y *few-shot*. Específicamente en el dominio médico, Agrawal et al. (2022) confirmaron que estos modelos operan como etiquetadores de alta eficacia para la extracción de información clínica sin requerir procesos de *fine-tuning* exhaustivos. En el dominio radiológico, Adams et al. (2023) utilizaron modelos como `GPT-4` para estructurar informes de texto libre y detectar hallazgos patológicos con alta precisión.

Bhayana et al. (2024); Le Guellec et al. (2024) advierten que la adopción directa de LLM comerciales en flujos de trabajo clínicos reales presenta barreras de privacidad al transmitir datos protegidos de pacientes a servicios de terceros, además de altos costos financieros y computacionales de inferencia masiva. En respuesta a esta restricción, autores como Mukherjee et al. (2023) evalúan el despliegue de modelos de pesos abiertos ejecutados localmente, como `Vicuna-13B`, para mantener la privacidad de los datos. Estos modelos locales continúan exigiendo infraestructuras de hardware escasas en la mayoría de los entornos clínicos estándar.

3.3 Destilación de conocimiento y modelos estudiantes ligeros

Como introducen Hinton et al. (2015), la destilación de conocimiento permite utilizar un modelo de gran escala (maestro) para generar etiquetas sintéticas destinadas al entrenamiento de modelos más eficientes (estudiantes). Gu et al. (2025) aplicaron este esquema en CheX-GPT y entrenaron un modelo ligero sobre 50.000 radiografías de tórax a partir de etiquetas de la familia GPT. Este método superó a los sistemas basados en reglas y a la inferencia directa del LLM maestro.

En el dominio de la TC de tórax, De Ferrari et al. (2025) demostraron que modelos de lenguaje de menor escala (`BETO` o `mBERT`), entrenados bajo supervisión débil con etiquetas generadas por modelos GPT, alcanzan consistencia inter-modelo ($\kappa \geq 0{,}78$) y superan estadísticamente a arquitecturas homólogas entrenadas con conjuntos limitados de datos revisados por expertos (F1-macro de 0.88 frente a 0.60). Estos resultados indican que la destilación de un oráculo LLM proporciona una señal de entrenamiento superior a la escasez de la anotación humana en el dominio de procesamiento de lenguaje natural.

3.4 Limitaciones computacionales en radiología 3D

La obtención de etiquetas textuales precisas mediante destilación de LLM no resuelve directamente la transferencia de esta señal de supervisión hacia clasificadores visuales tridimensionales. Si bien la alineación multimodal en radiología 3D utiliza modelos fundacionales como CT-CLIP para codificar volúmenes mediante aprendizaje contrastivo y *transformers* espaciales, su aplicación presenta barreras operativas. Específicamente, el entrenamiento *end-to-end* de arquitecturas volumétricas para acoplarlas a etiquetas débiles satura la memoria de video durante la retropropagación, lo que restringe su uso en entornos estándar.

Mientras que la literatura reciente prioriza la validación textual de la etiqueta clínica o propone arquitecturas fundacionales que asumen recursos de cómputo ilimitados para su ajuste, existe una necesidad de estrategias computacionalmente eficientes. Para abordar esta restricción, el presente trabajo desacopla el extractor visual del clasificador mediante el uso de perceptrones multicapa (MLP) sobre vectores visuales congelados (*frozen features*) extraídos por CT-CLIP, y se apoya en la validación del LLM como oráculo fiable. Bajo este esquema, el análisis de las leyes de escalamiento permite determinar si un alto volumen de etiquetas sintéticas débilmente supervisadas logra compensar los límites representacionales impuestos por un extractor visual estático.

4

Propuesta metodológica

La metodología consiste en un pipeline multimodal que desacopla la generación de etiquetas textuales del entrenamiento del clasificador visual. El flujo de trabajo utiliza el conjunto de datos CT-RATE y se divide en cuatro etapas: procesamiento de lenguaje natural para la anotación, extracción de características volumétricas, evaluación de cuellos de botella representacionales y estudio de leyes de escalamiento.

4.1 Formalización del problema y protocolo de datos

Sea $\mathcal{X}$ el dominio de volúmenes de TC 3D y $\mathcal{R}$ el dominio de informes radiológicos en texto libre. El problema se formula en un entorno de aprendizaje débilmente supervisado, compuesto por un conjunto de datos no etiquetado $\mathcal{D}_u = \{(\mathbf{X}_i, r_i)\}_{i=1}^{N}$ y un conjunto con etiquetas expertas $\mathcal{D}_e = \{(\mathbf{X}_j, \mathbf{y}_j)\}_{j=1}^{M}$, donde $M \ll N$ y el vector $\mathbf{y} \in \{0,1\}^C$ representa la presencia de C patologías torácicas objetivo.

4.1.1 Caracterización del conjunto de etiquetas expertas

El conjunto $\mathcal{D}_e$, utilizado como *gold standard* para la validación del oráculo y el entrenamiento de los modelos de referencia, proviene del subconjunto de informes del Hospital Universitario Medipol anotados manualmente por el equipo de CT-RATE. El proceso incluyó una fase de traducción y normalización supervisada por estudiantes avanzados de medicina.

Si bien la literatura original no reporta un índice de *inter-annotator agreement* para este proceso, estas anotaciones constituyen la referencia clínica oficial del corpus. Para efectos de esta investigación, se asume la validez de este conjunto como representación de la verdad de campo, y se reconoce como limitación la opacidad en los protocolos de consenso del conjunto original. No obstante, la coherencia de estas etiquetas se valida de forma indirecta en esta tesis mediante el análisis de estabilidad y alineación

semántica presentado en el Capítulo 5.

El objetivo es optimizar un clasificador visual $f_\theta : \mathcal{X} \to \{0,1\}^C$. Para mitigar la escasez de datos en $\mathcal{D}_e$ y el costo de la anotación manual, un LLM actúa como función de oráculo $\Lambda : \mathcal{R} \to \{0,1\}^C$. Mediante esta función, se generan etiquetas sintéticas $\hat{\mathbf{y}}_i = \Lambda(r_i)$ para la totalidad de $\mathcal{D}_u$, lo que permite entrenar los parámetros θ sobre los pares débilmente supervisados $\{(\mathbf{X}_i, \hat{\mathbf{y}}_i)\}_{i=1}^N$.

La validación experimental utiliza el conjunto de datos público CT-RATE Hamamci et al. (2026), desarrollado por la Istanbul Medipol University. El análisis se centra en cinco patologías torácicas: opacidad pulmonar, linfadenopatía, secuela fibrótica pulmonar, calcificación de la pared arterial y nódulo pulmonar. Dado el desbalance en la distribución de estas clases, se selecciona el F1-macro como métrica principal de evaluación.

Para comparar la utilidad de la supervisión débil frente a la supervisión experta, el corpus se divide en dos particiones disjuntas:

- **Subconjunto Manual (*gold standard*):** Contiene volúmenes con etiquetas verificadas por radiólogos ($\mathcal{D}_e$). Se utiliza para establecer las líneas base supervisadas bajo presupuestos compartidos (desde $N = 20$ hasta $N = 1520$) y suministra los *folds* de prueba para la validación cruzada.
- **Pool de Entrenamiento Débil:** Conformado por los volúmenes sin anotación experta ($\mathcal{D}_u$). Provee la base para la inferencia de etiquetas algorítmicas, lo que habilita la evaluación de regímenes de escalamiento asintótico hasta $N = 46438$.

4.2 Distribución de patologías y desbalance

A diferencia del *benchmark* CT-RATE original, que propone 18 categorías de anomalías, este trabajo restringe el análisis a las cinco patologías con mayor prevalencia en el *dataset*: calcificación de la pared arterial, linfadenopatía, nódulo pulmonar, opacidad pulmonar y secuela fibrótica pulmonar.

Esta decisión metodológica responde a dos criterios fundamentales:

1. **Densidad de datos:** El recorte garantiza un volumen de ejemplos positivos suficiente para observar la dinámica de las leyes de escalamiento. Esto evita que el sesgo de clase extrema en patologías infrecuentes actúe como un factor de confusión en la comparación entre etiquetas manuales y sintéticas.
2. **Representatividad clínica:** Estas clases cubren un espectro diverso de desafíos visuales —desde estructuras de alto contraste hasta hallazgos con alta varianza de textura—, lo que permite evaluar la capacidad de generalización del extractor de características estático Hamamci et al. (2026).

Un análisis de la prevalencia de clases en los conjuntos etiquetados, muestra que la supervisión débil preserva la distribución de las anotaciones expertas (ver Figura 4.1). El desbalance estructural del corpus se refleja en las proporciones de casos positivos (+) y negativos (-) (ver Tabla 4.1).

Tabla 4.1: Cuantificación del desbalance intra e interclase: recuentos absolutos de casos positivos (+) y negativos (-) en el subconjunto manual (N = 1520) y el conjunto anotado por `gpt-5-nano` (N = 46438).

Patología Objetivo	**Manual**		**GPT**	
	(+)	**(-)**	**(+)**	**(-)**
Nódulo pulmonar	621	899	20713	25725
Opacidad pulmonar	489	1031	18379	28059
Calcificación de la pared arterial	433	1087	13168	33270
Linfadenopatía	444	1076	11869	34569
Secuela fibrótica pulmonar	441	1079	10460	35978
Ninguna (0)	298	1222	8247	38191

La prevalencia varía entre patologías; por ejemplo, el nódulo pulmonar presenta cerca del doble de ocurrencias que la secuela fibrótica pulmonar. Esta heterogeneidad requiere el uso de métricas promediadas a nivel macro (F1-macro), de modo que todas las clases contribuyan de forma equitativa al rendimiento global evaluado.

Existe un desbalance intraclase, donde los casos negativos constituyen la clase mayoritaria en cada etiqueta (ver Tabla 4.1). Esta característica justifica el uso del AUPRC como métrica primaria para la selección de modelos, dado que permite evaluar el rendimiento sobre clases positivas minoritarias. Asimismo, motiva la incorporación de funciones de pérdida ponderadas durante el entrenamiento para mitigar el sesgo hacia la clase mayoritaria.

4.3 Estrategias de supervisión y generación de etiquetas

El diseño experimental utiliza dos fuentes de supervisión para el entrenamiento de los clasificadores visuales (manual y GPT). RadBERT, el codificador textual de CT-CLIP, se emplea exclusivamente en la validación del proceso de etiquetado.

- **Etiquetas Manuales (Línea Base):** Entrenamiento realizado sobre las anotaciones de $\mathcal{D}_e$. Representa el aprendizaje con datos verificados por expertos, cuyo volumen está limitado por el costo de

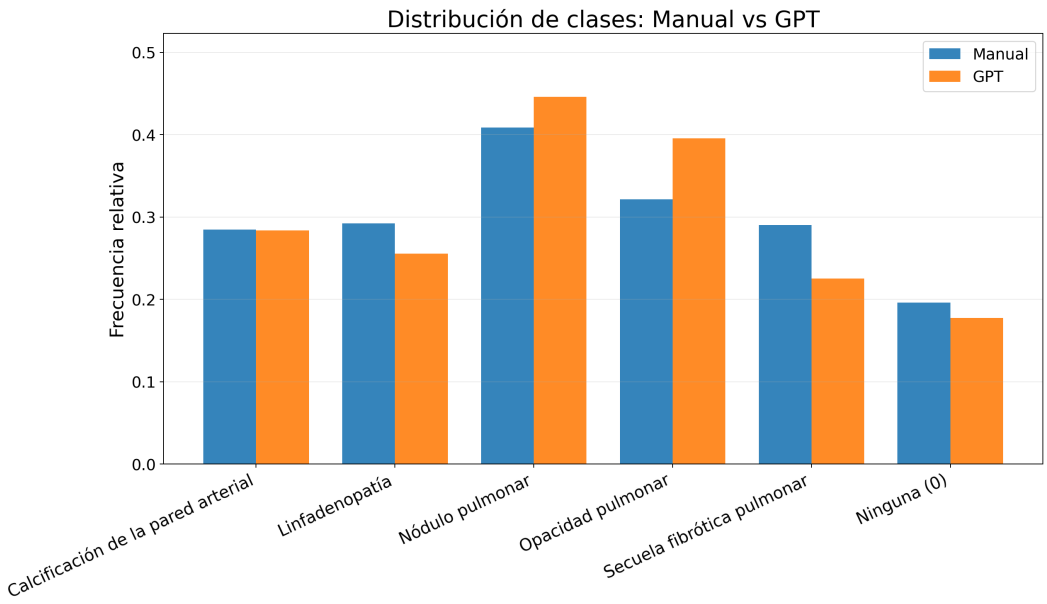


Figura 4.1: Distribución de frecuencia relativa de las patologías objetivo. Se contrasta la proporción de clases en el subconjunto manual frente al conjunto etiquetado por `gpt-5-nano`. El oráculo textual mantiene la distribución del *gold standard*.

anotación manual.

- **Etiquetas RadBERT (Referencia):** Generadas mediante el codificador textual de CT-CLIP, un modelo específico del dominio médico. Actúa como mecanismo de verificación secundaria para cuantificar la fidelidad de la extracción de etiquetas del LLM antes del entrenamiento de los modelos visuales.
- **Etiquetas GPT (Propuesta):** Utiliza un modelo de la familia GPT para generar anotaciones sobre $\mathcal{D}_u$. La configuración final emplea una estrategia *few-shot* (3-shot con restricciones negativas) diseñada para reducir falsos positivos en opacidad pulmonar y mejorar la sensibilidad en linfadenopatía. Se seleccionó la variante `gpt-5-nano` por su equilibrio entre rendimiento y costo computacional frente a modelos de mayor escala. Evaluaciones de estabilidad sobre el corpus masivo justifican el uso de una pasada de inferencia única (*single-pass*) para la generación del conjunto final de entrenamiento (ver Anexo B).

4.4 Representación visual y extracción de *features*

El sistema estudiado no entrena el codificador visual *end-to-end*, sino que usa representaciones precomputadas. El sistema opera sobre vectores latentes extraídos mediante el codificador visual de CT-

CLIP, tal como se ilustra en la Figura 4.2.

Este codificador emplea bloques de *transformers* 3D factorizados para abordar la complejidad cuadrática de la auto-atención tridimensional, la cual resulta prohibitiva en volúmenes médicos de alta resolución. La arquitectura factorizada desacopla el cálculo de la atención, lo que permite procesar la información espacial (intra-corte) de manera secuencial o independiente respecto a la información axial (inter-corte o profundidad). Esta factorización reduce el costo de memoria y de todos modos permite capturar dependencias entre características volumétricas.

La preparación de las imágenes incluyó una estandarización radiológica y espacial estricta. En el dominio espacial, los volúmenes originales se remuestrearon a una resolución física común de $0{,}75 \times 0{,}75 \times 1{,}5$ mm y se sometieron a una operación de recorte central y relleno (*padding*) para fijar las dimensiones del tensor de entrada en $480 \times 480 \times 240$ vóxeles. Las intensidades radiológicas se restringieron al rango de -1000 a 1000 unidades Hounsfield (HU) y se escalaron por un factor de $1/1000$ para garantizar la estabilidad numérica de la red.

Tras el preprocesamiento, el codificador transforma los tensores en representaciones estáticas dentro del espacio latente $\mathbb{R}^{512}$ y sus pesos se mantienen congelados durante todo el entrenamiento posterior. Para fundamentar empíricamente la viabilidad de este diseño frente al cuello de botella representacional, se evalúa el alineamiento semántico de estos vectores de forma previa al entrenamiento de los clasificadores. Este control de calidad preliminar cuantifica la similitud del coseno emparejada frente a la no emparejada y calcula las métricas de recuperación multimodal (MRR, MAP y *recall*), lo cual asegura que el espacio latente posee la capacidad intrínseca de discriminar estructuras patológicas de interés (los resultados de esta validación se detallan al inicio del Capítulo 5).

4.5 Diseño experimental y fases de entrenamiento

El entrenamiento de los perceptrones multicapa (MLP) sobre las representaciones latentes se organiza en un diseño de cuatro fases. Este diseño evalúa la capacidad de las representaciones, aísla variables de confusión y asegura una comparación estadísticamente robusta entre las fuentes de etiquetas.

1. **Fase 0 (Optimización de hiperparámetros):** Se utiliza el marco de optimización Optuna para identificar configuraciones estables (tasa de aprendizaje, decaimiento de pesos, dimensiones ocultas y *dropout*) de manera independiente para cada fuente de supervisión. Una vez identificada la mejor configuración, esta se congela y el modelo se entrena sobre cinco semillas aleatorias independientes. Este procedimiento cuantifica la varianza introducida por la inicialización de pesos y la partición de datos.

2. **Fase 1 (Evaluación de cuellos de botella):** Se entrena un clasificador estrictamente lineal (*linear probe*) sobre las *frozen features* para determinar si los límites de rendimiento provienen de la capacidad de la red o de la separabilidad de los vectores visuales. Adicionalmente, se ejecuta una optimización de umbrales de decisión sobre el conjunto de validación para maximizar el F1-macro en el conjunto de prueba.

3. **Fase 2 (Control de sensibilidad del protocolo):** Se implementa un diseño exploratorio de curvas de aprendizaje bajo el protocolo (*fixed-split*). El objetivo de esta fase no radica en establecer el rendimiento comparativo final, sino en diagnosticar empíricamente la sensibilidad de los clasificadores frente a los sesgos muestrales de los conjuntos estáticos. La cuantificación de esta distorsión justifica metodológicamente la implementación de la Fase 3.

4. **Fase 3 (Inferencia estadística y validación cruzada):** Constituye la evaluación confirmatoria para mitigar la varianza y establecer el punto de cruce (*crossover*). Se emplea el protocolo (*5-fold cross-validation*) con 5 semillas independientes por *fold*, para un total de 25 ejecuciones por cada presupuesto de datos. La evaluación abarca desde presupuestos compartidos (N = 20 a N = 1520) hasta regímenes asintóticos exclusivos para el etiquetado automático (hasta N = 46438).

   La significancia de las diferencias de rendimiento se analiza mediante la métrica delta emparejada (Manual frente a GPT) por cada par *fold*-semilla. Se calcula un intervalo de confianza *bootstrap* del 95 % para cada delta y se aplica una prueba de permutación emparejada. Para controlar comparaciones múltiples, se utiliza el método de Benjamini-Hochberg para el FDR, con un umbral de significancia de $p < 0{,}05$.

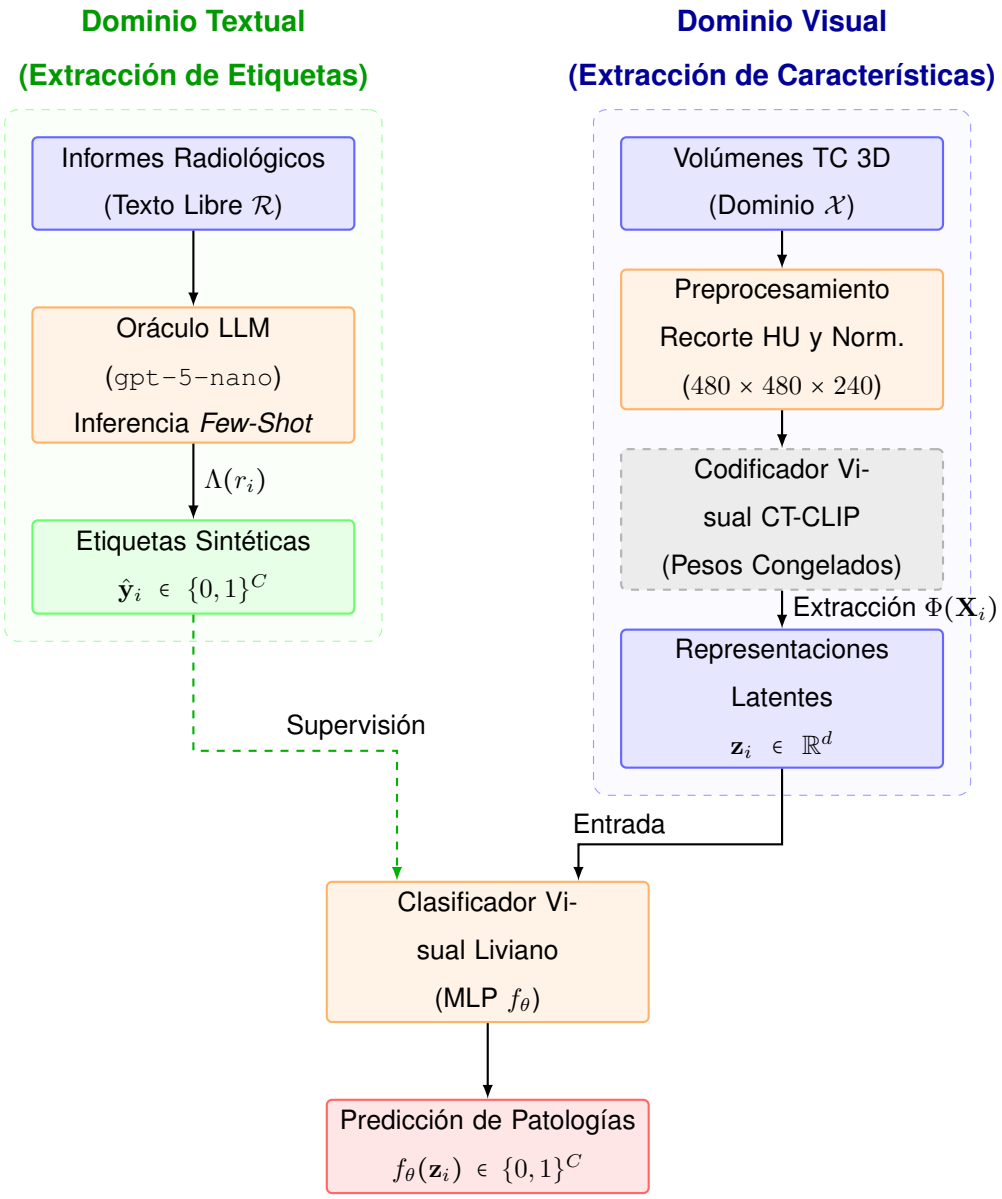


Figura 4.2: Arquitectura del pipeline multimodal. Se utiliza supervisión débil mediante etiquetas generadas por LLM para entrenar un clasificador sobre características visuales congeladas.

5

Resultados

5.1 Validación del espacio latente visual

Antes de evaluar los clasificadores y el oráculo textual, se cuantificó el alineamiento de las características volumétricas extraídas por el modelo CT-CLIP frente a sus respectivas narrativas clínicas. Sobre una muestra de 22185 pares de volumen e informe, la similitud del coseno media emparejada a nivel de instancia alcanzó un valor de $0{,}624$, frente a una similitud no emparejada de $-0{,}001$. A nivel semántico, la similitud media emparejada se situó en $0{,}183$, superando la media no emparejada de $-0{,}013$. Este margen de separación, respaldado por métricas de recuperación multimodal como MRR, MAP y *recall* (ver Tabla A.1 en el Anexo A), confirma empíricamente que el espacio latente posee una densidad semántica estructurada. La extracción visual captura las anomalías radiológicas sin requerir *fine-tuning* de la arquitectura base y establece una base geométrica válida para el entrenamiento de los perceptrones multicapa.

5.2 Rendimiento y estabilidad del oráculo textual

Los pacientes del conjunto de datos presentan frecuentemente multi-morbilidad, lo que requiere que los clasificadores operen bajo un esquema multi-etiqueta. Para verificar que las etiquetas sintéticas preservan las asociaciones patológicas presentes en el *gold standard*, se computó la correlación de Pearson entre las clases.

Al calcular la correlación de Pearson entre las 5 etiquetas, tanto en el conjunto manual y como el de GPT (ver Figura 5.1), se observa una diferencia absoluta media de $0{,}056$ entre los pares correspondientes de cada conjunto, lo que indica una alta similitud en las etiquetas asignadas. Asociaciones presentes en

las etiquetas manuales, como la relación entre linfadenopatía y calcificación arterial (r = 0,25) o entre linfadenopatía y opacidad pulmonar (r = 0,18), se mantienen en el dominio sintético (r = 0,22 y r = 0,17, respectivamente). La desviación máxima observada es de 0,11, lo que indica que la supervisión débil no introduce combinaciones de hallazgos clínicamente anómalas.

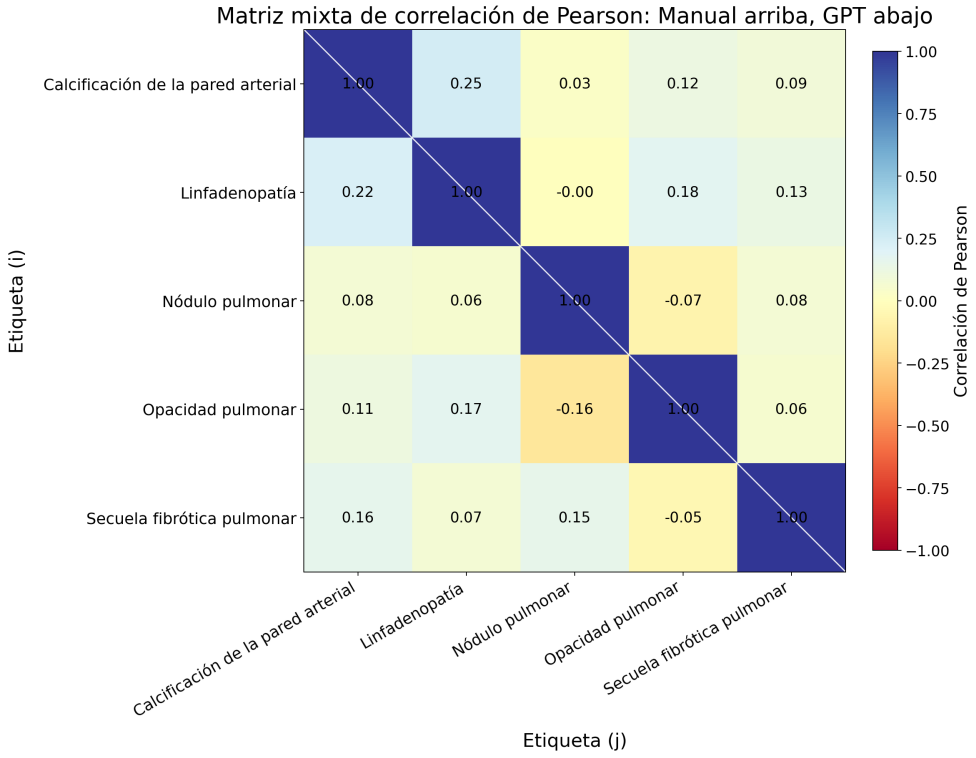


Figura 5.1: Matriz mixta de correlación de Pearson para las cinco patologías objetivo. En el triángulo superior se presentan las correlaciones del *gold standard* manual y el inferior las etiquetas generadas por `gpt-5-nano`. La similitud evidencia la preservación de las dependencias clínicas en el dominio sintético.

5.3 Generación de etiquetas mediante LLM

La configuración de prompts para el oráculo textual se definió mediante un proceso de optimización sobre un conjunto de ajuste reducido (N = 100). Tras evaluar distintas estrategias, se seleccionó un esquema *few-shot* con tres ejemplos y restricciones negativas explícitas aplicadas sobre el modelo `gpt-5-nano`. Esta configuración alcanzó un F1-macro promedio de 0,8934 ± 0,0078.

La validación final de la configuración seleccionada sobre el conjunto completo de test (N = 615) reportó un F1-macro de 0,8889 ± 0,0017.

En la ejecución de tres pasadas con `gpt-5-nano` sobre el conjunto de datos se observó una alta estabilidad. El *agreement* por pares registró un índice Kappa de Cohen promedio de $0{,}947$, con un acuerdo exacto a tres vías superior al $96\,\%$ en todas las clases (ver Figura 5.2).

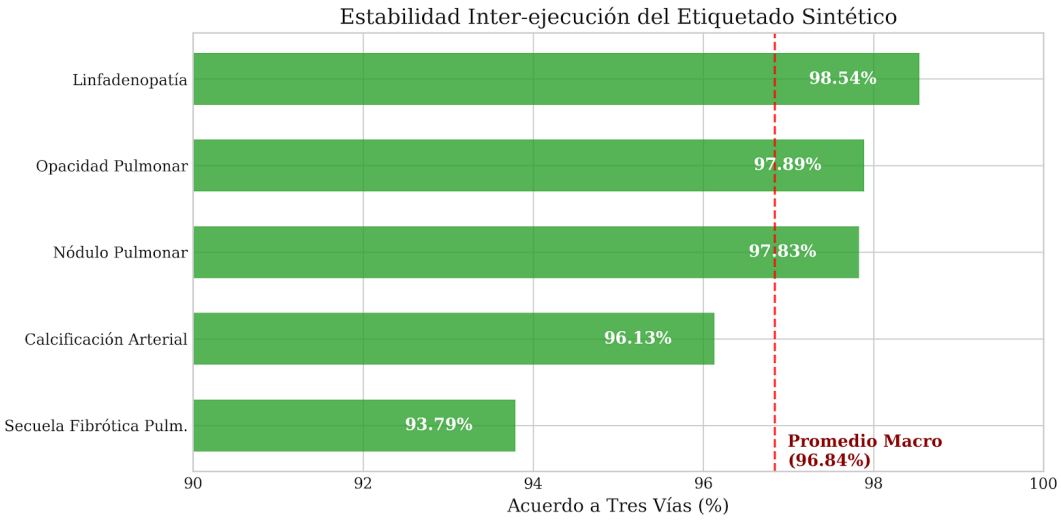


Figura 5.2: *Agreement* a tres vías para las etiquetas generadas por `gpt-5-nano` durante el etiquetado masivo, lo que muestra una estabilidad superior al $96\,\%$ en todas las clases evaluadas.

La evaluación de estrategias de ensamblaje por votación indicó que una regla de consenso basada en (any positive) incrementaba el F1-macro frente al *gold standard* a $0{,}9034$, lo que representó una ganancia absoluta marginal de $\sim 1{,}5\,\%$ respecto a una pasada individual. Dado este margen, se optó por utilizar las etiquetas de una inferencia *single-pass* para el entrenamiento de los clasificadores visuales, lo que reduce la complejidad computacional del pipeline sin afectar la fidelidad de la supervisión.

5.4 Extracción de características y restricciones representacionales

La evaluación del alineamiento semántico y la separabilidad de las características se realizó sobre los vectores latentes extraídos mediante los bloques de transformers 3D factorizados de CT-CLIP. La Fase 1 del diseño experimental implementó un clasificador lineal estricto (*linear probe*) como control de capacidad para determinar si los límites de rendimiento provenían de la arquitectura del clasificador o de la naturaleza de los vectores visuales.

Bajo un presupuesto compartido de $N = 1191$, el uso de un MLP frente a un clasificador lineal genera un incremento absoluto de $+0{,}0729$ en el AUPRC para los modelos entrenados con etiquetas manuales. Por el contrario, esta mejora se restringe a $+0{,}0229$ bajo supervisión sintética (ver Tabla 5.1).

Esta asimetría empírica evidencia que las etiquetas expertas logran capitalizar relaciones no lineales

Tabla 5.1: Comparativa de rendimiento entre clasificador lineal (*linear probe*) y MLP ($N = 1191$) sobre el conjunto de prueba fijo.

Fuente	**Modelo**	**AUPRC**	**AUROC**	**F1-macro**
Manual	Lineal	$0{,}5821 \pm 0{,}0185$	$0{,}7440 \pm 0{,}0152$	$0{,}5980 \pm 0{,}0151$
	MLP	$0{,}6550 \pm 0{,}0991$	$0{,}7746 \pm 0{,}0652$	$0{,}6181 \pm 0{,}0538$
GPT	Lineal	$0{,}5522 \pm 0{,}0077$	$0{,}7067 \pm 0{,}0037$	$0{,}5638 \pm 0{,}0146$
	MLP	$0{,}5751 \pm 0{,}0090$	$0{,}7157 \pm 0{,}0044$	$0{,}5934 \pm 0{,}0061$

en el espacio latente. La convergencia de métricas en la supervisión sintética confirma que el cabezal multicapa no extrae mayor señal discriminativa de las *frozen features* frente a un límite lineal estricto. Incluso al escalar la supervisión sintética a su límite máximo ($N = 46438$), el AUPRC se estabiliza en $0{,}566 \pm 0{,}028$, valor cercano al obtenido por el clasificador lineal con una fracción de los datos.

Se evaluó la optimización de umbrales de decisión para descartar que una selección subóptima penalizara métricas como el F1-macro. La optimización empírica por clase sobre el conjunto de validación incrementó el rendimiento absoluto en la evaluación de prueba (particularmente para los modelos entrenados con etiquetas sintéticas, cuyo F1-macro aumentó de $0{,}5237$ a $0{,}5934$). Este ajuste aisló el cuello de botella de optimización, pero no modificó el orden relativo de rendimiento entre las fuentes de etiquetas bajo este diseño de partición fija.

Estos controles de capacidad lineal y calibración, ejecutados durante la fase exploratoria sobre un conjunto de prueba estático, evidencian que el límite del sistema radica en las representaciones visuales congeladas. Si bien la evaluación bajo particiones fijas acarrea limitaciones inherentes de sesgo muestral —las cuales se mitigan mediante validación cruzada en la Sección 5.6—, la incapacidad de la red para superar el techo representacional mediante aumentos en la complejidad del cabezal fundamenta empíricamente la presencia de un cuello de botella visual transversal a ambas fuentes de etiquetas.

5.5 Sensibilidad del protocolo y diagnóstico de sesgo muestral

El rendimiento comparativo entre las fuentes de etiquetas demostró una alta sensibilidad respecto al protocolo de partición de datos. En la Fase 2, obtenidos bajo un esquema exploratorio de partición estática para el conjunto de prueba, se observó una falsa superioridad asintótica de la supervisión manual frente a la sintética.

El análisis de varianza inter-semilla determinó que este esquema de evaluación singular amplificaba el sesgo introducido por la selección aleatoria de los datos de entrenamiento y validación. Las fluctuaciones

marginales en las fronteras de decisión sesgaron las métricas macro-promediadas en el conjunto estático, lo que distorsionó la evaluación real de la eficiencia muestral entre las fuentes de etiquetas.

La identificación empírica de este sesgo invalida la partición estática como mecanismo para contrastar esquemas de supervisión débil y establece que la Fase 3, sustentada en validación cruzada y evaluación emparejada, constituye la única evidencia inferencial válida para determinar el punto de cruce entre ambas fuentes.

5.6 Entrenamiento de clasificadores y leyes de escalamiento

Utilizando las configuraciones arquitectónicas óptimas establecidas durante la exploración de hiperparámetros (Fase 0), se procedió a la construcción de las curvas de escalamiento bajo el marco inferencial previamente justificado de la Fase 3 (ver Figura 5.3). Las pruebas de estabilidad iniciales sobre estos hiperparámetros mostraron que los modelos entrenados con etiquetas manuales presentaban mayor varianza entre semillas que sus contrapartes entrenadas con etiquetas sintéticas, tendencia que se consolida en la evaluación general (ver Tabla 5.2).

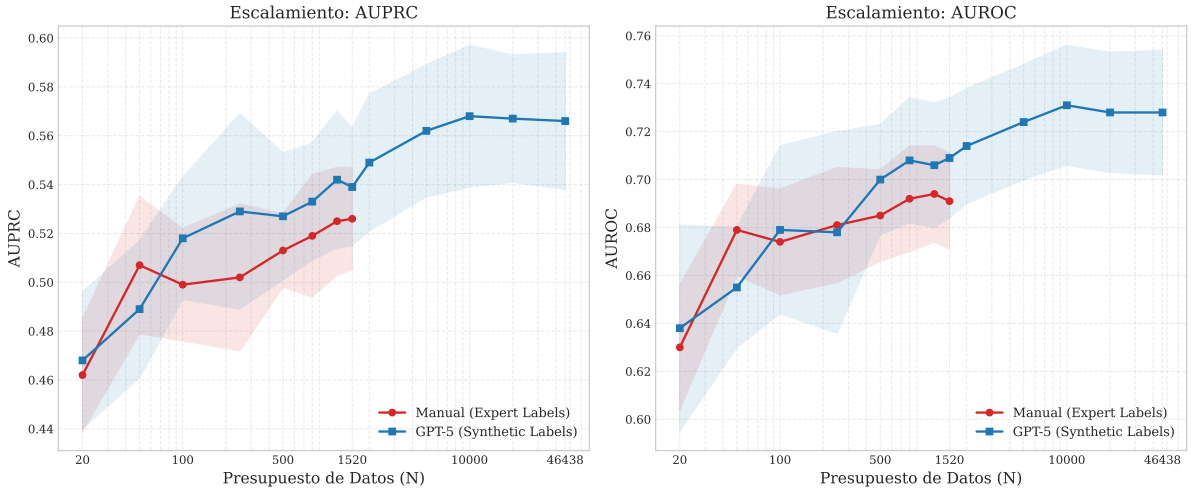


Figura 5.3: Curvas de escalamiento comparativas entre la supervisión manual y la generada por LLM. El AUPRC (izquierda) y el AUROC (derecha) se evalúan en función del presupuesto de datos N (eje horizontal en escala logarítmica). Los presupuestos bajos concentran la zona de volatilidad, previa a la consolidación empírica donde la escalabilidad sintética compensa a las anotaciones manuales.

Las pruebas de permutación emparejada con ajuste de FDR sobre las 25 ejecuciones por presupuesto (ver Anexo C) contradicen la premisa de que las etiquetas manuales poseen una eficiencia muestral

Tabla 5.2: Comparación inferencial del rendimiento de clasificadores (Fase 3). Los valores resaltados indican superioridad estadística ($p < 0{,}05$).

N	AUPRC		AUROC		F1-Macro	
	Manual	**GPT**	**Manual**	**GPT**	**Manual**	**GPT**
50	0,507 ± 0,028	0,489 ± 0,028	0,679 ± 0,019	0,655 ± 0,025	**0,513 ± 0,056**	0,472 ± 0,055
100	0,499 ± 0,023	**0,518 ± 0,025**	0,674 ± 0,022	**0,679 ± 0,035**	0,463 ± 0,080	**0,512 ± 0,046**
250	0,502 ± 0,030	**0,529 ± 0,040**	0,681 ± 0,024	0,678 ± 0,042	0,514 ± 0,035	0,516 ± 0,045
500	0,513 ± 0,015	**0,527 ± 0,026**	0,685 ± 0,019	**0,700 ± 0,023**	0,535 ± 0,023	0,540 ± 0,027
800	0,519 ± 0,025	**0,533 ± 0,024**	0,692 ± 0,022	**0,708 ± 0,026**	0,538 ± 0,026	**0,556 ± 0,025**
1191	0,525 ± 0,022	**0,542 ± 0,028**	0,694 ± 0,020	**0,706 ± 0,026**	0,534 ± 0,021	**0,549 ± 0,029**
1520	0,526 ± 0,021	**0,539 ± 0,024**	0,691 ± 0,020	**0,709 ± 0,025**	0,533 ± 0,020	**0,552 ± 0,027**
Rendimiento asintótico exclusivo (Sólo GPT)						
2000	– –	0,549 ± 0,028	– –	0,714 ± 0,024	– –	0,556 ± 0,021
5000	– –	0,562 ± 0,027	– –	0,724 ± 0,024	– –	0,565 ± 0,026
10000	– –	0,568 ± 0,029	– –	0,731 ± 0,025	– –	0,574 ± 0,023
20000	– –	0,567 ± 0,026	– –	0,728 ± 0,025	– –	0,575 ± 0,026
46438	– –	0,566 ± 0,028	– –	0,728 ± 0,026	– –	0,573 ± 0,026

universalmente superior. En el presupuesto mínimo (N = 20), la supervisión sintética presenta ventaja estadística en F1-Macro (Δ = −0,099, p_{adj} = 0,013), resultado atribuible a la alta varianza muestral característica de este régimen extremo. La supervisión manual superó estadísticamente a la sintética únicamente en N = 50, donde registró un AUROC de $0{,}679 \pm 0{,}019$ frente a $0{,}655 \pm 0{,}025$ de GPT (diferencia emparejada de +0,025, p_{adj} = 0,013) y un F1-Macro de $0{,}513 \pm 0{,}056$ frente a $0{,}472 \pm 0{,}055$ (Δ = +0,041, p_{adj} = 0,014).

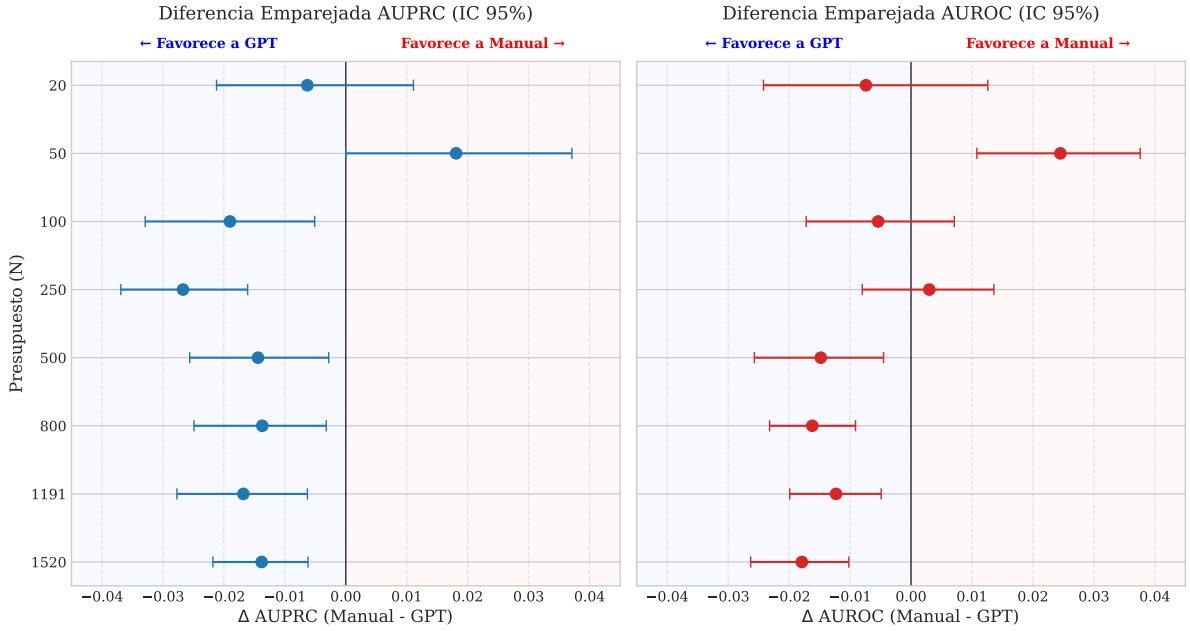


Figura 5.4: Evolución de la diferencia emparejada (Δ = Manual − GPT) y sus intervalos de confianza del 95 % a través de los presupuestos compartidos. Los intervalos que no cruzan la línea vertical de cero indican una diferencia estadísticamente significativa de acuerdo a la prueba de permutación.

A partir del presupuesto N = 100, las etiquetas generadas por GPT igualaron el rendimiento manual (por ejemplo, al alcanzar un AUROC de $0{,}679 \pm 0{,}035$ frente al $0{,}674 \pm 0{,}022$ manual). Desde la marca de N = 500, la supervisión sintética mostró una ventaja estadística consistente; en dicho punto, el modelo GPT registró un AUROC de $0{,}700 \pm 0{,}023$ comparado con el $0{,}685 \pm 0{,}019$ del modelo experto (diferencia emparejada de Δ = −0,015, p_{adj} = 0,015). En el presupuesto compartido máximo (N = 1520), los modelos GPT superaron a los manuales en todas las métricas principales y lograron un AUPRC de $0{,}539 \pm 0{,}024$ frente a $0{,}526 \pm 0{,}021$ (Δ = −0,013, p_{adj} = 0,013) y un AUROC de $0{,}709 \pm 0{,}025$ frente a $0{,}691 \pm 0{,}020$ (Δ = −0,018, p_{adj} = 0,002).

Al analizar por clase con N = 1520, se evidencia que la supervisión sintética generaliza de manera

consistente a través de distintas patologías, sin limitarse a un hallazgo específico. Por ejemplo, el modelo entrenado con etiquetas del LLM supera a su contraparte manual tanto en anomalías de alto contraste radiológico, como la calcificación de la pared arterial (F1-score de $0{,}718 \pm 0{,}055$ frente a $0{,}706 \pm 0{,}041$), como en patrones texturales complejos como la secuela fibrótica pulmonar (F1-score de $0{,}431 \pm 0{,}039$ frente a $0{,}393 \pm 0{,}053$).

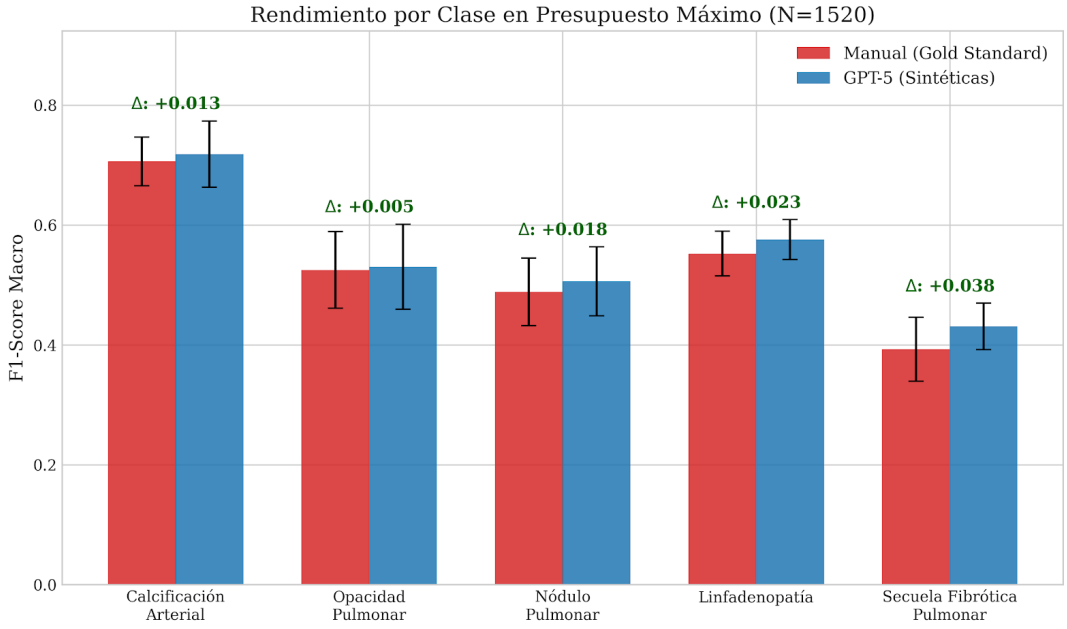


Figura 5.5: Análisis de rendimiento por clase (F1-Score) en el presupuesto compartido máximo ($N = 1520$).

La evaluación de la escalabilidad asintótica exclusiva de la supervisión débil se extendió hasta un presupuesto máximo de $N = 46438$. En los resultados se observa una fase de estancamiento de las métricas posterior a los $N = 10000$ ejemplos. Los modelos entrenados con $N = 10000$ etiquetas sintéticas alcanzaron el rendimiento máximo global en AUPRC ($0{,}568 \pm 0{,}029$) y AUROC ($0{,}731 \pm 0{,}025$). Al incrementar el presupuesto a $N = 20000$, se observó el valor máximo en F1-Macro ($0{,}575 \pm 0{,}026$), mientras que el uso de la totalidad del conjunto de datos ($N = 46438$) mantuvo el rendimiento general estabilizado dentro de los márgenes de error.

Este comportamiento asintótico indica que el aumento del volumen de datos compensa el ruido del etiquetado automático, en línea con la Hipótesis 2 sobre la viabilidad del escalamiento con LLM, y es consistente con la restricción impuesta por el cuello de botella de las características visuales precomputadas.

6

Conclusiones y trabajo futuro

Este trabajo de investigación evaluó la viabilidad de la supervisión débil basada en LLM para la clasificación de patologías en volúmenes de tomografía computarizada de tórax bajo restricciones computacionales. El diseño metodológico permitió cumplir los objetivos específicos planteados: se validó la consistencia del oráculo textual, se aisló el cuello de botella representacional mediante clasificadores lineales y se caracterizaron las leyes de escalamiento de ambas fuentes de supervisión.

El análisis experimental indica que las etiquetas generadas mediante `gpt-5-nano` constituyen una fuente de supervisión competitiva frente a la anotación experta. Aunque las anotaciones manuales presentan ventajas en regímenes de datos muy reducidos, la supervisión sintética iguala y supera su rendimiento a partir de presupuestos moderados ($N \geq 100$) y mantiene una ventaja estadísticamente significativa en métricas como AUROC y AUPRC en regímenes superiores.

El rendimiento comparativo demostró ser sensible al protocolo de partición de datos. El uso de particiones fijas (*fixed-splits*) amplifica los sesgos muestrales y distorsiona la evaluación entre fuentes de etiquetas. La implementación de un marco inferencial basado en validación cruzada acoplada a múltiples semillas de inicialización permitió modelar esta incertidumbre y evidenció empíricamente la superioridad de la supervisión sintética. Esta divergencia advierte sobre el riesgo metodológico de utilizar particiones estáticas al evaluar esquemas de supervisión débil.

El análisis estadístico de las curvas de escalamiento descarta la existencia de una eficiencia muestral universal para la supervisión manual. En regímenes de datos iniciales ($N \leq 50$), el rendimiento relativo exhibe una alta volatilidad, con ventajas alternantes dependientes de la semilla aleatoria. No obstante, al superar este umbral de inestabilidad, el aumento en el volumen de datos compensa el ruido del etiquetado automático, lo que permite a la supervisión sintética consolidar una ventaja estadísticamente significativa frente a su contraparte experta.

A pesar de esta ventaja, el rendimiento global se estabiliza de forma asintótica en presupuestos altos. Este estancamiento confirma que operar exclusivamente con representaciones visuales estáticas de CT-CLIP genera un cuello de botella.

En conclusión, estos hallazgos establecen que la supervisión débil mediante LLM es una estrategia viable y eficiente para entornos con limitaciones de anotación experta y recursos computacionales, aunque sujeta al límite representacional que imponen las características visuales congeladas. Las principales limitaciones de este estudio radican en la dependencia de una única arquitectura fundacional de extracción visual y la restricción de hardware que impidió la retropropagación de gradientes hacia el codificador.

Como trabajo futuro, se propone explorar esquemas que permitan relajar esta restricción representacional mediante estrategias de *fine-tuning* parcial del codificador visual, o bien, a través de arquitecturas que integren de manera más estrecha la información textual y visual tridimensional sin incurrir en costos computacionales prohibitivos.

Bibliografía

Adams, L. C., Truhn, D., Busch, F., Kader, A., Niehues, S. M., Makowski, M. R., and Bressem, K. K. (2023). Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: A multilingual feasibility study. *Radiology*, 307(4):e230725.

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. (2022). Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1998–2022.

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Bhayana, R. et al. (2024). Chatbots and large language models in radiology: what the radiologist needs to know. *Radiology*, 310(1):e232756.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Buzug, T. M. (2008). *Computed tomography: from photon statistics to medical image management*. Springer Science & Business Media.

Chng, S. Y., Tern, P. J. W., Kan, M. R. X., and Cheng, L. T. E. (2023). Automated labelling of radiology reports using natural language processing: Comparison of traditional and newer methods. *Health Care Science*, 2(2):64–75.

De Ferrari, J., Ñanculef, R., Benoit, D., Araya, M., and Solar, M. (2025). Assessing GPT as a weak oracle for annotating radiological studies. page 98–109, Berlin, Heidelberg. Springer-Verlag.

Diaz-Pinto, A., Alle, S., Nath, V., Tang, Y., Ihsani, A., Asad, M., Pérez-García, F., et al. (2024). MONAI label: A framework for AI-assisted interactive labeling of 3D medical images. *Medical Image Analysis*, 95:103207.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Draelos, R. L., Dover, D., MacEachern, M. J., Rakshit, S., Berman, R., Viglianti, B. L., Carin, L., Duke, K., and Mazurowski, M. A. (2021). Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical Image Analysis*, 67:101857.

Dunnmon, J. A., Ratner, A. J., Saab, K., Rubin, N. B., Ré, C., and Rubin, D. L. (2020). Cross-modal data programming enables rapid medical machine learning. *Patterns*, 1(2):100019.

Frénay, B. and Verleysen, M. (2013). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT Press.

Gu, J., You, K., Cho, H.-C., Kim, J., Hong, E. K., and Roh, B. (2025). CheX-GPT: Harnessing large language models for enhanced chest X-ray report labeling. *Radiology*, 314(3):e241476.

Hamamci, I. E., Er, S., Hou, E. S., et al. (2026). Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*. Publicado originalmente como arXiv:2403.17834.

Hansell, D. M., Bankier, A. A., MacMahon, H., McLoud, T. C., Müller, N. L., and Remy, J. (2008). Fleischner society: Glossary of terms for thoracic imaging. *Radiology*, 246(3):697–722.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hounsfield, G. N. (1973). Computerized transverse axial scanning (tomography): Part 1. description of system. *The British Journal of Radiology*, 46(552):1016–1022.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciari, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Le Guellec, B. et al. (2024). Performance of an open-source large language model in extracting information from free-text radiology reports. *Radiology: Artificial Intelligence*, 6(3):e230267.

Lipton, Z. C., Elkan, C., and Naryanaswamy, B. (2014). Optimal thresholding of classifiers to maximize f1 measure. In Calders, T., Esposito, F., Hüllermeier, E., and Meo, R., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 225–239, Berlin, Heidelberg. Springer Berlin Heidelberg.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530.

Mukherjee, P., Hou, B., Lanfredi, R. B., and Summers, R. M. (2023). Feasibility of using the privacy-preserving large language model Vicuna for labeling radiology reports. *Radiology*, 309(1):e231147.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282.

Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A. Y., and Lungren, M. P. (2020). Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1519.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Yan, A., Sanghavi, H., Xu, Y., et al. (2022). RadBERT: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53.

Anexos

A Validación de alineamiento semántico multimodal

Se evaluó el alineamiento del espacio latente visual durante el control de calidad preliminar. La evaluación se ejecutó sobre una muestra empírica de 22185 pares de volumen tomográfico e informe radiológico, procesados mediante la arquitectura CT-CLIP con pesos congelados. Para cuantificar el desempeño, se calculó el MRR, MAP y textitrecall bajo los enfoques de recuperación de visual a texto y de texto a visual (ver Tabla A.1).

Tabla A.1: Métricas de recuperación a nivel de instancia y semántico para el espacio latente compartido.

Nivel	**Dirección**	**MRR**	**MAP**	**Recall@1**	**Recall@5**	**Recall@10**	**Recall@50**
Instancia	Visual a Texto	0.0156	0.0184	0.0079	0.0203	0.0314	0.0965
	Texto a Visual	0.0172	0.0151	0.0075	0.0215	0.0378	0.1081
Semántico	Visual a Texto	0.2145	0.1990	0.1418	0.2952	0.3970	0.7173
	Texto a Visual	0.2847	0.2102	0.1638	0.4132	0.5603	0.8710

Los resultados a nivel de instancia y semántico aportan evidencia complementaria sobre la estructura clínica del espacio latente. En el nivel de instancia, la tarea exige identificar un par exacto entre 22185 alternativas; bajo esta estricta restricción, las métricas obtenidas superan holgadamente la probabilidad aleatoria y evidencian un anclaje multimodal efectivo. A su vez, la evaluación semántica amplía el criterio de éxito hacia la coincidencia de etiquetas patológicas, donde el *recall*@50 alcanza 0.8710 en la dirección de texto a visual. Esta progresión verifica que la arquitectura proyecta las características visuales y textuales en vecindades geométricas agrupadas por anomalías compartidas. La proximidad vectorial responde a la presencia de patologías y no solo a la memorización de pares aislados, lo que fundamenta el uso de estas representaciones congeladas para entrenar los perceptrones multicapa.

B Optimización de prompts y selección de oráculo textual

La fase inicial de ingeniería de prompts evaluó distintas configuraciones sobre un conjunto de ajuste reducido (N = 100). La aproximación base (*zero-shot*) presentó limitaciones asociadas a la generación de alucinaciones, particularmente en la clase de opacidad pulmonar, donde registró una precisión de $0{,}5540 \pm 0{,}0142$ y una baja sensibilidad para linfadenopatía ($0{,}3793 \pm 0{,}0488$). La transición a un esquema *few-shot* con tres ejemplos (3-shot multi v3) y restricciones negativas explícitas corrigió estas deficiencias. Esta configuración elevó el F1-macro promedio a $0{,}8934 \pm 0{,}0078$ y maximizó la sensibilidad en linfadenopatía ($1{,}0$) en la muestra evaluada.

Tabla B.1: Evolución del rendimiento durante la optimización del prompt y selección de modelo evaluado sobre el conjunto de ajuste reducido (N = 100).

Modelo	**Configuración**	**F1-Macro**
`gpt-5-nano`	Zero-shot	$0{,}7895 \pm 0{,}0109$
`gpt-5-nano`	3-shot v1	$0{,}7677 \pm 0{,}0069$
`gpt-5-nano`	3-shot v2	$0{,}8346 \pm 0{,}0043$
`gpt-5-nano`	3-shot v3	$0{,}8934 \pm 0{,}0078$
`gpt-5-mini`	3-shot v3	$0{,}9123 \pm 0{,}0029$
`gpt-5.1`	3-shot v3	$0{,}9180 \pm 0{,}0017$

La evaluación de escalabilidad del modelo de lenguaje indicó que el uso de arquitecturas de mayor capacidad (`gpt-5-mini` y `gpt-5.1`) proporcionó incrementos marginales en el F1-macro ($+0{,}0189$ y $+0{,}0246$, respectivamente, frente a `gpt-5-nano`). Estos márgenes de mejora no justificaron el incremento sustancial en los costos computacionales asociados a la inferencia (incrementos de un factor de 5 y 25, respectivamente), por lo que se seleccionó `gpt-5-nano` para el etiquetado masivo del corpus completo.

C Resultados detallados de inferencia estadística

El análisis de las leyes de escalamiento de la Sección 5.6 generó los valores de inferencia estadística emparejada (ver Tabla C.1). La evaluación midió las diferencias de rendimiento (Δ = Manual – GPT) a través de los presupuestos compartidos de entrenamiento y empleó una prueba de permutación emparejada. El método de Benjamini-Hochberg (FDR) controló la tasa de falsos descubrimientos derivados de las comparaciones múltiples. Un valor $p_{adj} < 0{,}05$ señala la superioridad estadística de una fuente de etiquetas frente a la otra para un presupuesto y métrica dados.

Tabla C.1: Resultados de la inferencia estadística emparejada (Δ = Manual – GPT) sobre los presupuestos compartidos (Fase 3). Se reporta la diferencia media, el intervalo de confianza (IC) del $95\,\%$ y el valor p ajustado por FDR. En los valores p en negrita se indican las diferencias estadísticamente significativas.

Métrica	Presupuesto (N)	Δ Media	IC 95 %	Valor p_{adj} (FDR)
AUPRC	20	−0,0063	[−0,0212, +0,0111]	0,5255
	50	+0,0181	[−0,0000, +0,0371]	0,0977
	100	−0,0190	[−0,0329, −0,0051]	**0,0254**
	250	−0,0267	[−0,0369, −0,0161]	**0,0024**
	500	−0,0144	[−0,0256, −0,0028]	**0,0330**
	800	−0,0137	[−0,0249, −0,0032]	**0,0330**
	1191	−0,0168	[−0,0277, −0,0063]	**0,0140**
	1520	−0,0138	[−0,0218, −0,0062]	**0,0132**
AUROC	20	−0,0074	[−0,0242, +0,0126]	0,5255
	50	+0,0245	[+0,0108, +0,0376]	**0,0134**
	100	−0,0054	[−0,0172, +0,0071]	0,5182
	250	+0,0030	[−0,0080, +0,0136]	0,6390
	500	−0,0148	[−0,0257, −0,0045]	**0,0153**
	800	−0,0162	[−0,0232, −0,0091]	**0,0024**
	1191	−0,0123	[−0,0199, −0,0049]	**0,0134**
	1520	−0,0179	[−0,0263, −0,0102]	**0,0024**
F1-Macro	20	−0,0989	[−0,1566, −0,0444]	**0,0134**
	50	+0,0412	[+0,0154, +0,0663]	**0,0140**
	100	−0,0497	[−0,0834, −0,0193]	**0,0134**
	250	−0,0021	[−0,0245, +0,0191]	0,8580
	500	−0,0044	[−0,0172, +0,0083]	0,5704
	800	−0,0182	[−0,0305, −0,0061]	**0,0171**
	1191	−0,0153	[−0,0245, −0,0061]	**0,0134**
	1520	−0,0194	[−0,0316, −0,0075]	**0,0134**