

PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE CIENCIAS EXACTAS NATURALES Y AMBIENTALES

CARRERA DE CIENCIA DE DATOS

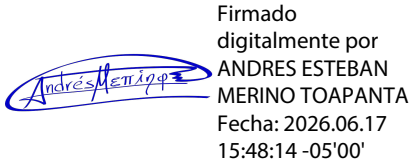
Técnicas de fine-tuning para Large Language Models: revisión sistemática y evaluación comparativa

Monografía previa a la obtención del título de Ingeniero en Ciencia de Datos

Francisco Javier Flores Sánchez

Quito, 2026

Certifico que la Monografía de Ciencia de Datos de **Francisco Javier Flores Sánchez** ha sido concluida de conformidad con las normas establecidas; por lo tanto, puede ser presentada para la calificación correspondiente.



Andrés Esteban Merino Toapanta

Tutor de la monografía

Quito, 25 de junio de 2026

DEDICATORIA

A mi familia y a quienes conforman mi círculo cercano,

por creer en mí y en esta carrera,

mucho antes de que este proyecto existiera.

AGRADECIMIENTOS

A Andrés Merino, por haber sido un apoyo que trasciende el rol de tutor y profesor. Su amistad, orientación y cercanía en este mundo de la ciencia de datos, disciplina aún emergente en el país, marcaron profundamente mi formación. Agradezco su disposición constante, su paciencia y la confianza que depositó en mí y en este trabajo desde las primeras etapas de la carrera hasta la culminación de este proyecto.

A Bernardo Villegas, por ser un excelente primer jefe y, sobre todo, un amigo que me mostró que la ciencia de datos va mucho más allá de lo que se aprende en el aula. Su visión me ayudó a comprender la grandeza de esta carrera y su conexión con la inteligencia artificial y sus múltiples derivados. Le agradezco profundamente por guiarme en la elección del tema, por creer en este proyecto desde su concepción.

TABLA DE CONTENIDOS

1. Resumen **1**

2. Abstract **2**

3. Introducción **3**

4. Objetivos **6**

4.1. Objetivo general 6

4.2. Objetivos específicos 6

5. Fundamento teórico **7**

5.1. Grandes modelos de lenguaje y la arquitectura Transformer 7

5.2. Técnicas de *fine-tuning* 8

5.2.1. *Full Fine-Tuning* y *Supervised Fine-Tuning* 8

5.2.2. Métodos PEFT: LoRA, QLoRA, *Adapter Layers*, *Prefix Tuning* y *Prompt Tuning* 10

5.2.3. Métodos de alineación: RLHF y DPO 13

5.3. Métricas de evaluación 14

5.3.1. Métricas de clasificación 14

5.3.2. Métricas de generación de texto 15

6. Metodología **17**

6.1. Estrategia de búsqueda 17

6.2. Proceso de selección de estudios 18

6.2.1. Fase 1: Cribado de título y resumen 19

6.2.2. Fase 2: Evaluación de texto completo 20

6.3. Evaluación de calidad metodológica 22

6.4. Clasificación por dominio de aplicación 22

6.5. Extracción de datos 23

6.6. Análisis de datos 24

7. Resultados y discusión **25**

7.1. Caracterización del corpus 25

7.2. Resultados por dominio 26
7.2.1. Salud y Biomedicina 26
7.2.2. Lenguaje y Humanidades 29
7.2.3. Ciencias Exactas y Computación 30
7.2.4. Industria y Negocios 32
7.3. Análisis comparativo entre dominios 34
7.3.1. Distribución de técnicas 34
7.3.2. Cobertura metodológica y reporte de métricas 36
7.3.3. Patrones transversales 36

8. Conclusiones **39**

9. Referencias **42**

10. Apéndices **54**

Apéndice A: Corpus completo de estudios incluidos 54
A.1 Salud y Biomedicina 54
A.2 Lenguaje y Humanidades 56
A.3 Ciencias Exactas y Computación 57
A.4 Industria y Negocios 59

1. RESUMEN

Esta monografía presenta una revisión sistemática de las técnicas de *fine-tuning* para *Large Language Models* (LLMs) publicadas entre 2020 y 2025, conducida mediante el protocolo PRISMA 2020. A partir de una búsqueda inicial de 140 registros en Scopus, se seleccionaron 84 estudios empíricos con evaluación cuantitativa, distribuidos en cuatro dominios: Salud y Biomedicina (35,7 %), Lenguaje y Humanidades (25,0 %), Ciencias Exactas y Computación (23,8 %) e Industria y Negocios (15,5 %). El análisis abarcó nueve técnicas organizadas en tres paradigmas: *Full Fine-Tuning* y SFT como enfoques de actualización completa de parámetros, cinco métodos de *Parameter-Efficient Fine-Tuning* (LoRA, QLoRA, *Adapter Layers*, *Prefix Tuning*, *Prompt Tuning*) y dos métodos de alineación (RLHF, DPO). Los resultados evidencian la hegemonía de LoRA y sus variantes, presente en el 79,8 % de los estudios incluidos, con una calidad metodológica del corpus aceptable (*score* promedio de 2,61 sobre 3,0). Se identificaron tres patrones transversales: la efectividad consistente del *fine-tuning* sobre el modelo base sin adaptar, la ventaja de las técnicas PEFT en lenguas y dominios de baja cobertura en el preentrenamiento, y la tendencia hacia arquitecturas compuestas en aplicaciones de producción. Los hallazgos permiten formular criterios de selección basados en evidencia empírica según el contexto operativo, las restricciones de privacidad y los requisitos de alineación.

Palabras clave: *fine-tuning*, *Large Language Models*, LoRA, *Parameter-Efficient Fine-Tuning*, RLHF, DPO, revisión sistemática, PRISMA, procesamiento de lenguaje natural, evaluación empírica.

2. ABSTRACT

This monograph presents a systematic review of fine-tuning techniques for Large Language Models (LLMs) published between 2020 and 2025, conducted following the PRISMA 2020 protocol. From an initial search of 140 records in Scopus, 84 empirical studies with quantitative evaluation were selected, distributed across four application domains: Health and Biomedicine (35.7 %), Language and Humanities (25.0 %), Exact Sciences and Computing (23.8 %), and Industry and Business (15.5 %). The analysis covered nine techniques organized into three paradigms: Full Fine-Tuning and SFT as full-parameter update approaches, five Parameter-Efficient Fine-Tuning methods (LoRA, QLoRA, Adapter Layers, Prefix Tuning, Prompt Tuning), and two alignment methods (RLHF, DPO). Results reveal the dominance of LoRA and its variants, present in 79.8 % of included studies, with an acceptable methodological quality of the corpus (average score of 2.61 out of 3.0). Three cross-domain patterns were identified: the consistent effectiveness of fine-tuning over the unadapted base model, the advantage of PEFT techniques in low-resource languages and underrepresented domains, and the trend toward composite architectures in production applications. The findings enable evidence-based technique selection criteria according to operational context, data privacy constraints, and response alignment requirements.

Keywords: fine-tuning, Large Language Models, LoRA, Parameter-Efficient Fine-Tuning, RLHF, DPO, systematic review, PRISMA, natural language processing, empirical evaluation.

3. INTRODUCCIÓN

En los últimos años, los grandes modelos de lenguaje (*Large Language Models*, LLMs) han transformado múltiples áreas mediante avances en procesamiento de lenguaje natural (*Natural Language Processing*, NLP). Modelos como GPT-4, LLaMA y Mistral han demostrado capacidades emergentes en tareas de comprensión y generación de lenguaje, permitiendo aplicaciones que van desde asistentes conversacionales hasta análisis de documentos especializados y generación de código.

El rendimiento de estos modelos no depende únicamente de su arquitectura o datos de pre-entrenamiento, sino también de las técnicas de *fine-tuning* (ajuste fino) empleadas para adaptarlos a dominios específicos. El paradigma de pre-entrenar y ajustar modelos de lenguaje a tareas específicas se ha establecido como el enfoque estándar en el área (Ding et al., 2023, p. 220), permitiendo que modelos generales se especialicen mediante entrenamiento adicional en datos de dominio. Este principio de especialización se fundamenta en evidencia empírica que demuestra que los modelos adaptados mediante *fine-tuning* pueden capturar mejor el conocimiento específico de la tarea, conduciendo a rendimiento superior en dominios restringidos comparado con modelos de propósito general.

Desde 2020 han surgido múltiples técnicas representativas de *fine-tuning*, cada una con enfoques distintos de adaptación. En el marco de esta revisión, se analizan nueve técnicas representativas organizadas en tres familias según su paradigma: *Full Fine-Tuning* y su variante *Supervised Fine-Tuning* (SFT), que representan el enfoque tradicional de actualización completa de parámetros sobre datos etiquetados; los métodos de *Parameter-Efficient Fine-Tuning* (PEFT), como LoRA, QLoRA, *Adapter Layers*, *Prefix Tuning* y *Prompt Tuning*, que priorizan eficiencia reduciendo drásticamente el número de parámetros entrenables; y los métodos de alineación (*Alignment Methods*), como RLHF y DPO, que ajustan el comportamiento del modelo para reflejar preferencias y valores humanos. La coexistencia de estos tres paradigmas en la literatura reciente plantea preguntas sobre cuándo cada enfoque resulta más adecuado según el contexto operativo, pregunta central que esta revisión busca responder con base en evidencia empírica.

La evidencia empírica sobre estas técnicas se encuentra dispersa en múltiples fuentes académicas, donde cada publicación reporta comparaciones o evaluaciones según objetivos específicos del estudio. La proliferación de métodos ha sido notable: en los últimos años se han publicado más de cien artículos sobre PEFT (Lialin et al., 2024, p. 2), reflejando el intenso interés de la comunidad investigadora. Algunos trabajos comparan técnicas dentro de un paradigma, otros evalúan una técnica en múltiples dominios de aplicación, mientras que otros se enfocan en dominios particulares como procesamiento de textos médicos, generación de código o análisis de documentos legales. Esta especialización natural, aunque valiosa para avanzar conocimiento en áreas específicas, genera evidencia fragmentada que dificulta la consolidación de conocimiento comparativo entre técnicas y entre dominios.

La comparabilidad directa de resultados entre estudios se ve limitada por heterogeneidad en diseños experimentales. Estudios comparativos recientes evidencian que incluso los artículos más representativos carecen de comparabilidad completa, debido a diferencias en configuraciones de evaluación como conteos de parámetros variables o ausencia de ciertas métricas (Lialin et al., 2024, p. 35). La variabilidad en resultados reportados es sustancial: estudios experimentales documentan amplios rangos de rendimiento según la combinación de técnica, escala de modelo y configuración experimental empleada (Lialin et al., 2024, p. 29, Tabla 4). Adicionalmente, la combinación óptima de técnicas PEFT "puede variar para diferentes *backbones* de PLM, tareas *downstream* y escalas de datos" (Ding et al., 2023, p. 221), demostrando que el rendimiento depende críticamente del contexto experimental específico. En ausencia de síntesis comparativa sistemática, la selección de técnicas frecuentemente se basa en disponibilidad de implementaciones públicas o recomendaciones anecdóticas, más que en evidencia empírica consolidada.

Una revisión sistemática empleando el protocolo PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) aborda estas limitaciones mediante un proceso estructurado y transparente que documenta explícitamente la estrategia de búsqueda, establece criterios claros de inclusión y exclusión, y registra sistemáticamente el proceso de selección y extracción de datos. Esta sistematicidad permite consolidar evidencia dispersa, identificar patrones recurrentes en rendimiento según dominios de aplicación y detectar vacíos en comparaciones existentes. La propia literatura reconoce esta necesidad: la falta de comparabilidad entre estudios y la ausencia de protocolos de evaluación unificados constituyen limitaciones recurrentes en

el campo (Lialin et al., 2024, p. 35). Desde la perspectiva de la Ciencia de Datos, esta revisión se enfoca en métricas de calidad de predicción y patrones de aplicabilidad según dominio, excluyendo deliberadamente el análisis de eficiencia computacional dado que la heterogeneidad de implementaciones y hardware impediría comparaciones válidas entre estudios.

El presente trabajo aborda la siguiente pregunta de investigación:

> ¿Cuáles son las características, el rendimiento reportado en distintos dominios de aplicación, y los patrones de aplicabilidad de las técnicas de *fine-tuning* para *Large Language Models* desarrolladas entre 2020 y 2025, según la evidencia reportada en la literatura académica?

El alcance de esta revisión se delimita de la siguiente manera. Se analizan nueve técnicas representativas que cubren los tres paradigmas identificados: *Full Fine-Tuning* y SFT como enfoques de actualización completa de parámetros, cinco técnicas de *Parameter-Efficient Fine-Tuning* (LoRA, QLoRA, *Adapter Layers*, *Prefix Tuning*, *Prompt Tuning*) y dos métodos de alineación (RLHF, DPO). La revisión cubre literatura publicada entre 2020 y 2025, período que corresponde a la explosión metodológica en técnicas de *fine-tuning* tras la introducción de modelos de lenguaje de gran escala. Se incluyen artículos de revistas académicas con evaluación empírica cuantitativa, publicados en inglés y con acceso abierto, asegurando reproducibilidad del proceso de selección. El análisis se organiza en torno a cuatro dominios de aplicación: Salud y Biomedicina, Lenguaje y Humanidades, Ciencias Exactas y Computación, e Industria y Negocios. La aplicación de estos criterios mediante metodología PRISMA resultó en la inclusión de 84 estudios empíricos, cuya evidencia constituye la base del análisis comparativo presentado en esta monografía.

4. OBJETIVOS

El presente capítulo define el objetivo general y los objetivos específicos que orientan esta revisión sistemática, estableciendo el alcance y los productos esperados del trabajo.

4.1 OBJETIVO GENERAL

Comparar las técnicas de *fine-tuning* para *Large Language Models* mediante revisión sistemática PRISMA, identificando relaciones entre características técnicas, rendimiento en distintos dominios de aplicación, y patrones de aplicabilidad reportados en la literatura académica entre 2020 y 2025.

4.2 OBJETIVOS ESPECÍFICOS

Para alcanzar el objetivo general, se plantean tres objetivos específicos que cubren las etapas de sistematización, análisis y síntesis de la evidencia empírica.

- Sistematizar la literatura sobre *fine-tuning* de *Large Language Models* publicada entre 2020 y 2025 aplicando metodología PRISMA, extrayendo datos sobre características técnicas, configuraciones experimentales, métricas de evaluación y resultados de rendimiento en distintos dominios de aplicación.
- Analizar el rendimiento de las técnicas de *fine-tuning* según dominio de aplicación, estableciendo *trade-offs* y patrones de aplicabilidad reportados en la literatura.
- Sintetizar criterios de selección de técnicas según características del contexto operativo (disponibilidad de datos, dominio de aplicación, objetivos de rendimiento), identificando limitaciones reportadas y direcciones futuras de investigación.

5. FUNDAMENTO TEÓRICO

El presente capítulo establece el marco conceptual necesario para comprender las técnicas de *fine-tuning* de *Large Language Models*. Se inicia con una descripción de los LLMs y la arquitectura Transformer que los sustenta, seguida del paradigma de pre-entrenamiento y adaptación que motiva el *fine-tuning*. A continuación se presenta una taxonomía de las nueve técnicas representativas analizadas en esta revisión. Posteriormente se define el marco de evaluación empleado en la literatura, incluyendo métricas de rendimiento y *benchmarks* estándar.

5.1 GRANDES MODELOS DE LENGUAJE Y LA ARQUITECTURA TRANSFORMER

Los *Large Language Models* son modelos de lenguaje basados en la arquitectura Transformer, que se ha establecido como el fundamento de los avances recientes en procesamiento de lenguaje natural. El bloque constitutivo central del Transformer consiste en una capa de atención multi-cabeza (*multi-head attention*, MHA) seguida de una capa completamente conectada (*feed-forward network*, FFN), ambas con conexiones residuales y normalización de capas para mejorar la capacidad de entrenamiento (Lialin et al., 2024, p. 3). El mecanismo de atención calcula un promedio ponderado normalizado de los *tokens* de entrada, donde los pesos de atención se computan mediante el producto punto entre vectores de consulta (*query*) y clave (*key*) (Lialin et al., 2024, p. 3). Cada capa Transformer contiene dos sub-capas principales: la capa de atención y la capa *feedforward*, ambas seguidas de una proyección hacia la dimensión original de entrada y una conexión residual (Houlsby et al., 2019, p. 3).

Sobre esta arquitectura, los LLMs se entrenan mediante el paradigma de pre-entrenamiento y adaptación. En la fase de pre-entrenamiento, el modelo se entrena sobre corpus masivos de texto no etiquetado mediante objetivos auto-supervisados, adquiriendo representaciones lingüísticas generales que luego se transfieren a aplicaciones específicas (Ding et al., 2023, p. 220). La escala de estos modelos ha crecido de forma sostenida: los PLMs modernos emplean transformers con arquitecturas bidireccionales, autoregresivas o de secuencia a secuencia, entrenados sobre corpus de escala masiva (Ding et al., 2023, p. 220). Esta diversidad arquitectónica y de escala es precisamente lo que hace necesario un análisis sistemático de las técnicas de adaptación, ya que

el comportamiento del fine-tuning varía según la arquitectura base y el dominio de aplicación.

Sin embargo, el pre-entrenamiento por sí solo no es suficiente para obtener el comportamiento deseado en aplicaciones específicas. Un paradigma central en NLP consiste en el pre-entrenamiento a gran escala sobre datos de dominio general seguido de la adaptación a tareas o dominios particulares (Hu et al., 2021, p. 1). Esta adaptación, conocida como *fine-tuning*, constituye el objeto de estudio de la presente revisión sistemática.

5.2 TÉCNICAS DE *FINE-TUNING*

Las técnicas de *fine-tuning* se organizan en tres familias según su paradigma de adaptación, taxonomía que estructura tanto el marco teórico de esta revisión como la organización del corpus analizado. La primera familia es ***Full Fine-Tuning***, el enfoque tradicional que actualiza todos los parámetros del modelo pre-entrenado; en esta revisión se analiza también *Supervised Fine-Tuning* (SFT), variante que aplica esta actualización completa sobre datos de instrucciones etiquetadas. La segunda familia es ***Parameter-Efficient Fine-Tuning*** (PEFT), que agrupa métodos diseñados para reducir drásticamente el número de parámetros entrenables manteniendo un rendimiento comparable al *full fine-tuning*; los cinco representantes analizados son LoRA, QLoRA, *Adapter Layers*, *Prefix Tuning* y *Prompt Tuning*. La tercera familia son los **métodos de alineación**, que ajustan el comportamiento del modelo según preferencias humanas en lugar de optimizar directamente sobre una tarea; se analizan RLHF y DPO. Estas nueve técnicas no agotan el espacio de métodos de adaptación existentes, sino que representan las de mayor presencia en la literatura empírica reciente, tal como refleja el corpus de 84 estudios analizado en esta revisión. A continuación se describe cada técnica con su formulación matemática y mecanismo de adaptación.

5.2.1 *FULL FINE-TUNING* Y *SUPERVISED FINE-TUNING*

FULL FINE-TUNING

Full Fine-Tuning constituye el enfoque tradicional de adaptación, donde todos los parámetros del modelo pre-entrenado se actualizan durante el entrenamiento sobre datos etiquetados

de la tarea objetivo. El objetivo de optimización consiste en maximizar la verosimilitud sobre el conjunto de entrenamiento:

$$\max_{\phi} \sum_{(x,y)\in\mathcal{D}} \log p_\phi(y \mid x), \tag{5.1}$$

donde ϕ representa el conjunto completo de parámetros del modelo, $\mathcal{D}$ es el dataset de entrenamiento, x son las entradas e y las salidas esperadas. Todos los pesos de la red se actualizan mediante gradiente descendente estocástico.

Este enfoque constituye el *baseline* (referencia de comparación) metodológico dado que maximiza la capacidad de adaptación del modelo. Sin embargo, a medida que los modelos crecen en escala, "hacer *fine-tuning* y almacenar todos los parámetros es prohibitivamente costoso y eventualmente se vuelve prácticamente inviable" (Ding et al., 2023, p. 220). Adicionalmente, para N tareas distintas el *full fine-tuning* requiere mantener N copias completas del modelo (Houlsby et al., 2019, p. 4), lo que representa un costo de almacenamiento proporcional al número de aplicaciones.

SUPERVISED FINE-TUNING

El *Supervised Fine-Tuning* (SFT) es una variante de *Full Fine-Tuning* que aplica la actualización completa de parámetros sobre datos de instrucciones etiquetadas en formato pregunta-respuesta. En el contexto del pipeline de alineación de modelos de lenguaje, SFT corresponde a la primera etapa: los anotadores humanos proporcionan demostraciones del comportamiento deseado y el modelo se ajusta sobre esos datos mediante aprendizaje supervisado (Ouyang et al., 2022, p. 6).

A diferencia del *fine-tuning* orientado a tareas específicas, el objetivo de SFT es que el modelo aprenda a seguir instrucciones en lenguaje natural. En la literatura reciente aparece tanto como técnica independiente, cuando el objetivo es mejorar la capacidad de seguir instrucciones sin pasos adicionales de alineación, como punto de partida para los métodos RLHF y DPO descritos más adelante en esta sección.

5.2.2 MÉTODOS PEFT: LORA, QLORA, *ADAPTER LAYERS*, *PREFIX TUNING* Y *PROMPT TUNING*

LORA (*LOW-RANK ADAPTATION*)

Low-Rank Adaptation (LoRA) parte de la hipótesis de que las actualizaciones de pesos durante la adaptación poseen un rango intrínseco bajo, lo que permite representarlas mediante una descomposición de matrices de rango reducido en lugar de actualizar directamente los pesos del modelo (Hu et al., 2021, p. 2). Sobre esta base, LoRA representa la actualización de una matriz de pesos $W_0 \in \mathbb{R}^{d\times k}$ como:

$$h = W_0 x + \Delta W x = W_0 x + BAx, \tag{5.2}$$

donde W_0 son los pesos pre-entrenados que permanecen congelados, $B \in \mathbb{R}^{d\times r}$ y $A \in \mathbb{R}^{r\times k}$ son las matrices entrenables, con $r \ll \text{mín}(d, k)$ siendo el rango de la descomposición (Hu et al., 2021, p. 4). La inicialización es estratégica: A se inicializa con distribución Gaussiana aleatoria y B con ceros, de modo que $\Delta W = BA$ es cero al inicio del entrenamiento, garantizando que el modelo comienza idéntico al pre-entrenado (Hu et al., 2021, p. 4).

En la fase de inferencia, las matrices B y A pueden fusionarse con W_0 para obtener $W = W_0 + BA$, sin introducir latencia adicional comparado con el modelo original (Hu et al., 2021, p. 2). Con rangos muy bajos, como uno o dos, LoRA resulta suficiente incluso cuando la dimensión completa del modelo es tan alta como 12,288, como en GPT-3 175B, lo que lo convierte en un método eficiente tanto en almacenamiento como en cómputo (Hu et al., 2021, p. 2).

QLORA (*QUANTIZED LOW-RANK ADAPTATION*)

Quantized Low-Rank Adaptation (QLoRA) extiende LoRA combinando la adaptación de bajo rango con cuantización de 4 bits del modelo base (Dettmers et al., 2023, p. 1). El método introduce tres innovaciones técnicas: cuantización NF4 (*Normal Float 4*), un tipo de datos de 4 bits óptimo para pesos distribuidos normalmente; doble cuantización (*double quantization*), que cuantiza las constantes de cuantización para reducir el uso de memoria; y optimizadores

paginados (*paged optimizers*), que gestionan los picos de memoria durante el entrenamiento (Dettmers et al., 2023, p. 2).

En QLoRA el modelo base se mantiene congelado en precisión de 4 bits mientras los adaptadores LoRA se entrenan en BFloat16. La operación de paso hacia adelante se expresa como:

$$Y^{BF16} = X^{BF16} \cdot \text{dequant}(W^{NF4}) + X^{BF16} \cdot B^{BF16}A^{BF16}, \tag{5.3}$$

donde W^{NF4} son los pesos del modelo base cuantizados a 4 bits, B y A son las matrices LoRA en BFloat16, y *dequant* denota la operación de descuantización para el cómputo (Dettmers et al., 2023, p. 5). Los gradientes se propagan únicamente a través de las matrices LoRA; los pesos cuantizados permanecen sin actualización.

ADAPTER LAYERS

Los módulos *adapter*, propuestos por Houlsby et al. (2019), son capas adicionales insertadas en la arquitectura Transformer que mantienen congelados los pesos pre-entrenados. Se insertan dos módulos *adapter* seriales en cada capa: uno tras la sub-capa de atención multi-cabeza y otro tras la sub-capa *feedforward* (Houlsby et al., 2019, p. 3).

La arquitectura de los *adapters* implementa un diseño de cuello de botella (*bottleneck*): proyectan las características originales de dimensión d a una dimensión menor m, aplican una no linealidad, y proyectan de vuelta a d dimensiones (Houlsby et al., 2019, p. 3):

$$h \leftarrow f(hW_{\text{down}})W_{\text{up}} + h, \tag{5.4}$$

donde $W_{\text{down}} \in \mathbb{R}^{d \times m}$ es la proyección descendente, $W_{\text{up}} \in \mathbb{R}^{m \times d}$ es la proyección ascendente, $f(\cdot)$ es una función de activación no lineal, y el término $+h$ representa la conexión residual interna (Houlsby et al., 2019, p. 3). En la práctica se utiliza entre el 0.5 y el 8 % de los parámetros del modelo original, y la inicialización cercana a cero garantiza que el módulo comienza como una función identidad aproximada, preservando el comportamiento del modelo pre-entrenado (Houlsby et al., 2019, p. 3).

PREFIX TUNING

Prefix Tuning mantiene congelados todos los parámetros del modelo y optimiza una secuencia de vectores continuos específicos de tarea, denominados *prefix*, que se anteponen a las activaciones de cada capa Transformer, permitiendo que los *tokens* subsecuentes atiendan al *prefix* como si fueran *virtual tokens* (X. L. Li y Liang, 2021, p. 4582). Para cada posición i en la secuencia, la activación se define como:

$$h_i = \begin{cases} P_\theta[i,:], & \text{si } i \in P_{idx} \\ \text{LM}_\phi(z_i, h_{<i}), & \text{en otro caso} \end{cases}, \tag{5.5}$$

donde P_θ es una matriz entrenable de dimensión $|P_{idx}| \times \dim(h_i)$, P_{idx} denota los índices del *prefix*, y LM_ϕ representa el modelo de lenguaje con parámetros congelados (X. L. Li y Liang, 2021, p. 4585). El *prefix* guía el comportamiento generativo del modelo sin modificar sus pesos, actuando como contexto continuo optimizable que influye tanto en la codificación de la entrada como en la distribución de los *tokens* generados.

En términos de eficiencia paramétrica, modificando solo el 0.1 % de los parámetros *Prefix Tuning* alcanza un rendimiento comparable al *full fine-tuning* en escenarios con datos completos y supera al *fine-tuning* en escenarios de datos limitados (X. L. Li y Liang, 2021, p. 4582).

PROMPT TUNING

Prompt Tuning (Lester et al., 2021, p. 3046) optimiza exclusivamente una secuencia de *tokens* virtuales continuos (*soft prompt*) que se anteponen a la entrada, manteniendo todos los parámetros del modelo congelados. A diferencia de *Prefix Tuning*, que inserta vectores entrenables en cada capa Transformer, *Prompt Tuning* opera únicamente en la capa de *embeddings* de entrada: los *tokens* entrenables se concatenan con los *tokens* de la tarea y el conjunto se procesa a través del modelo sin modificación:

$$h_0 = [P_\theta;\, E(x)], \tag{5.6}$$

donde $P_\theta \in \mathbb{R}^{l \times d}$ es la matriz de *prompt* entrenable de longitud l y dimensión d, $E(x)$ es la representación de *embedding* de la entrada x, y $[\cdot;\cdot]$ denota la concatenación (Lester et al., 2021,

p. 3047). Solo los parámetros de P_θ se actualizan durante el entrenamiento, representando una fracción mínima del total de parámetros del modelo.

Lester et al. (2021) demostraron que con modelos de escala suficiente, *Prompt Tuning* alcanza un rendimiento comparable al *full fine-tuning*, con la ventaja de que un único modelo base puede servir múltiples tareas simplemente intercambiando el *soft prompt* (Lester et al., 2021, p. 3046). La diferencia conceptual respecto a *Prefix Tuning* es que este último propaga vectores entrenables a través de todas las capas del Transformer, mientras que *Prompt Tuning* restringe la intervención a la capa de entrada.

5.2.3 MÉTODOS DE ALINEACIÓN: RLHF Y DPO

RLHF (*REINFORCEMENT LEARNING FROM HUMAN FEEDBACK*)

RLHF es una metodología para alinear modelos de lenguaje con preferencias humanas, abordando la discrepancia entre el objetivo de pre-entrenamiento y el comportamiento deseado en aplicaciones prácticas (Ouyang et al., 2022). El proceso se estructura en tres etapas secuenciales (Ouyang et al., 2022, p. 6): en la primera se recolectan demostraciones humanas del comportamiento deseado y se entrena una política supervisada; en la segunda se recopilan comparaciones entre pares de salidas del modelo donde evaluadores humanos indican sus preferencias, y se entrena un modelo de recompensa para predecir la salida preferida; en la tercera se optimiza la política usando la señal del modelo de recompensa mediante el algoritmo PPO (*Proximal Policy Optimization*), bajo el siguiente objetivo:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}[r(x,y)] - \beta\mathbb{D}_{\mathrm{KL}}[\pi_\theta(y|x)\|\pi_{\mathrm{ref}}(y|x)], \tag{5.7}$$

donde π_θ es la política optimizada, $r(x,y)$ es el modelo de recompensa, π_{ref} es la política de referencia tras el *Supervised Fine-Tuning*, y β controla la restricción de divergencia KL que previene que el modelo se desvíe excesivamente de la política de referencia (Rafailov et al., 2023, p. 3).

DPO (*DIRECT PREFERENCE OPTIMIZATION*)

DPO es una alternativa simplificada a RLHF que elimina la necesidad de entrenar un modelo de recompensa explícito y de emplear algoritmos de aprendizaje por refuerzo. El método utiliza un cambio de variables para definir la pérdida de preferencias directamente como función de la política, permitiendo optimizar un modelo de lenguaje desde datos de preferencias humanas mediante un objetivo simple de entropía cruzada binaria, sin muestrear de la política durante el entrenamiento (Rafailov et al., 2023, p. 2). El objetivo de optimización DPO se define como:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right], \tag{5.8}$$

donde σ es la función logística, (x, y_w, y_l) representa un triplete de *prompt*, respuesta preferida y respuesta no preferida, π_θ es la política optimizada, π_{ref} es la política de referencia, y β controla la restricción de divergencia KL implícita (Rafailov et al., 2023, p. 4). El gradiente de esta pérdida incrementa la verosimilitud de las completaciones preferidas y_w y disminuye la de las no preferidas y_l, ponderando cada ejemplo según qué tan incorrectamente el modelo de recompensa implícito ordena las completaciones (Rafailov et al., 2023, p. 4).

5.3 MÉTRICAS DE EVALUACIÓN

La evaluación del rendimiento de las técnicas de *fine-tuning* en la literatura emplea métricas específicas según el tipo de tarea evaluada. En esta revisión se presentan las métricas de mayor frecuencia en el corpus analizado, seleccionadas según la jerarquía de métricas empleada en la extracción de datos: para cada estudio se identificó la métrica principal declarada por los autores o, en su ausencia, la de mayor prioridad según el tipo de tarea. Las métricas se agrupan en dos categorías: clasificación y generación de texto.

5.3.1 MÉTRICAS DE CLASIFICACIÓN

Las métricas de clasificación evalúan la corrección de las predicciones discretas del modelo. Son especialmente frecuentes en tareas de procesamiento de lenguaje natural donde el objetivo es asignar una etiqueta discreta a una entrada de texto.

La ***accuracy*** (exactitud) mide la proporción de predicciones correctas sobre el total de instancias evaluadas, tomando valores entre 0 y 1, donde valores más altos indican mejor rendimiento. Constituye la métrica más directa para tareas de clasificación con clases balanceadas.

El **F1 Score** combina la *precision* (precisión) y el *recall* (recuperación) del modelo mediante su media armónica:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{5.9}$$

donde *precision* mide la proporción de predicciones positivas correctas y *recall* mide la proporción de positivos reales correctamente identificados (Chang et al., 2024, p. 27). El F1 Score varía entre 0 y 1, donde valores más altos indican mejor balance entre ambas métricas. Es especialmente útil cuando existe desbalance de clases, siendo frecuente en tareas de extracción de información y clasificación de texto especializado.

5.3.2 MÉTRICAS DE GENERACIÓN DE TEXTO

Las métricas de generación de texto evalúan la calidad del texto producido por el modelo mediante comparación con referencias humanas.

BLEU (*Bilingual Evaluation Understudy*) evalúa la calidad del texto generado mediante la precisión de n-gramas modificada entre el texto candidato y referencias humanas, combinada con una penalización por brevedad para evitar que textos cortos obtengan puntuaciones artificialmente altas (Papineni et al., 2002, p. 5). BLEU puede expresarse como proporción entre 0 y 1 o como porcentaje entre 0 y 100 según la implementación y las convenciones del estudio; en ambos casos valores más altos indican mayor solapamiento léxico con las referencias. En el corpus analizado en esta revisión ambas notaciones aparecen, por lo que los valores de BLEU se interpretan según la escala reportada por cada estudio.

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) adopta un enfoque orientado al *recall*: cuantifica la proporción de n-gramas de las referencias que aparecen en el texto candidato (C.-Y. Lin, 2004, p. 1). La variante ROUGE-L utiliza la subsecuencia común más larga (*Longest Common Subsequence*) entre el candidato y la referencia, capturando estructura

secuencial más allá de la co-ocurrencia simple de n-gramas (C.-Y. Lin, 2004, p. 2). ROUGE puede expresarse como proporción entre 0 y 1 o como porcentaje entre 0 y 100 según la implementación y las convenciones del estudio; en ambos casos valores más altos indican mayor solapamiento con las referencias humanas. En el corpus analizado en esta revisión ambas notaciones aparecen, por lo que los valores de ROUGE se interpretan según la escala reportada por cada estudio.

BERTScore supera las limitaciones de las métricas basadas en n-gramas al calcular la similitud semántica mediante *embeddings* contextuales de modelos pre-entrenados como BERT (T. Zhang et al., 2020, p. 1). Para una referencia x y un candidato $\hat{x}$, BERTScore computa precisión, *recall* y F1 mediante correspondencia greedy entre *tokens*:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, \quad F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}, \tag{5.10}$$

donde $\mathbf{x}_i$ y $\hat{\mathbf{x}}_j$ son los vectores de *embedding* contextual de cada *token* (T. Zhang et al., 2020, p. 4). BERTScore varía entre 0 y 1, donde valores más altos indican mayor similitud semántica con la referencia. A diferencia de BLEU y ROUGE, BERTScore puede identificar correspondencias semánticas entre palabras que difieren en forma superficial, como sinónimos o paráfrasis, lo que lo hace especialmente adecuado para evaluación en dominios especializados.

6. METODOLOGÍA

El presente capítulo describe el proceso metodológico empleado para identificar, seleccionar y analizar la literatura sobre técnicas de *fine-tuning* para *Large Language Models*. Se adopta el protocolo PRISMA 2020 (Page et al., 2021, p. 796), un estándar para la conducción y reporte de revisiones sistemáticas que organiza el proceso en cuatro fases secuenciales: identificación de registros mediante búsqueda en bases de datos, cribado por título y resumen, evaluación de elegibilidad en texto completo, e inclusión del corpus definitivo. La adopción de este protocolo garantiza que cada decisión de inclusión o exclusión quede documentada de forma explícita y reproducible.

6.1 ESTRATEGIA DE BÚSQUEDA

La estrategia de búsqueda se diseñó siguiendo el marco PICO adaptado a revisiones de NLP, donde la población corresponde a *Large Language Models*, la intervención a las técnicas de *fine-tuning* en scope, la comparación a estudios con evaluación cuantitativa, y el resultado a métricas de rendimiento reportadas. A partir de este marco se definieron cuatro grupos de términos de búsqueda: términos de adaptación (*fine-tuning*, *finetuning*, *fine tuning*), términos de modelo (*large language model*, *LLM*), términos de técnica (LoRA, *low-rank adaptation*, *adapter*, *prefix tuning*, *prompt tuning*, RLHF, DPO, *parameter-efficient*, PEFT) y términos de evaluación (*evaluation*, *benchmark*, *comparison*, *empirical*, *experimental*).

La búsqueda sistemática se realizó en la base de datos Scopus, seleccionada por su cobertura amplia de literatura académica en ciencias de la computación y procesamiento de lenguaje natural. La búsqueda se ejecutó sobre títulos, resúmenes y palabras clave mediante el siguiente *query string*:

```
TITLE-ABS-KEY(
  ("fine-tuning" OR "finetuning" OR "fine tuning")
  AND ("large language model" OR "large language models"
       OR "LLM" OR "LLMs")
  AND (
    "LoRA" OR "low-rank adaptation" OR
    "adapter" OR "adapters" OR
    "prefix tuning" OR "prompt tuning" OR
```

```
    "RLHF" OR "reinforcement learning from human feedback" OR
    "DPO" OR "direct preference optimization" OR
    "parameter-efficient" OR "PEFT"
  )
  AND (
    "evaluation" OR "benchmark" OR "benchmarking" OR
    "comparison" OR "comparative" OR
    "empirical" OR "experimental"
  )
)
AND PUBYEAR > 2019 AND PUBYEAR < 2026
AND DOCTYPE(ar)
AND OA(all)
AND LANGUAGE(english)
```

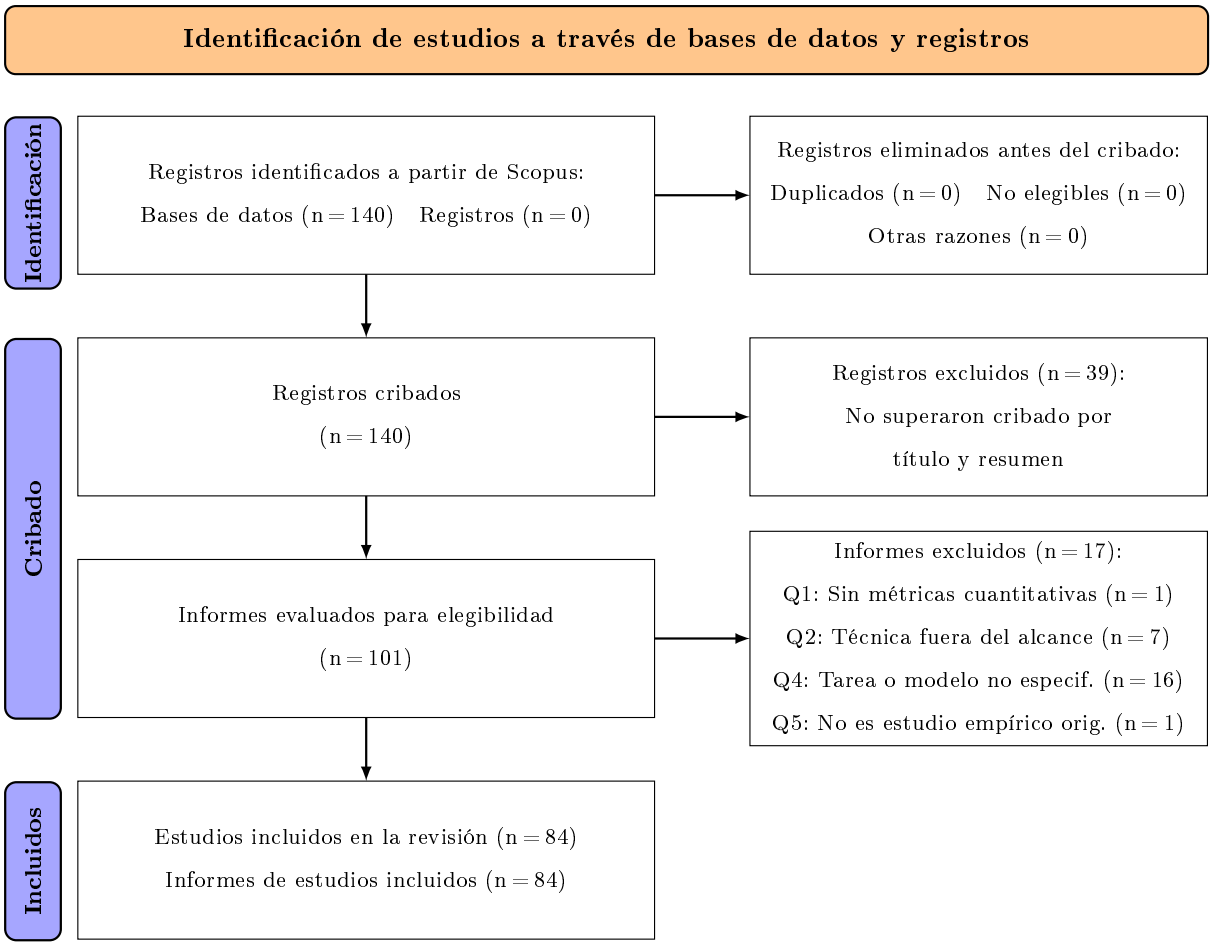
Se aplicaron los siguientes filtros directamente en la plataforma: período de publicación 2020-2025, tipo de documento restringido a artículos de revista, acceso abierto e idioma inglés. La búsqueda retornó 140 registros, los cuales fueron sometidos al proceso de selección descrito en la sección siguiente.

6.2 PROCESO DE SELECCIÓN DE ESTUDIOS

El proceso de selección se estructuró en dos fases secuenciales de cribado siguiendo el protocolo PRISMA 2020. La Figura 6.1 presenta el diagrama de flujo completo con los registros procesados en cada etapa. Todo el proceso fue implementado mediante código Python con llamadas estructuradas a APIs de modelos de lenguaje, garantizando reproducibilidad y trazabilidad de cada decisión.

Figura 6.1

Diagrama de flujo PRISMA 2020 del proceso de selección de estudios



Nota. Elaborado siguiendo las directrices de Page et al. (2021).

6.2.1 FASE 1: CRIBADO DE TÍTULO Y RESUMEN

Los 140 registros identificados en Scopus fueron evaluados mediante un sistema automatizado de votación por mayoría implementado con GPT-4o-mini. Cada registro recibió cuatro evaluaciones independientes, cada una con un perfil de evaluador distinto: experto en revisiones sistemáticas de NLP, revisor académico senior especializado en LLMs, investigador en aprendizaje automático, y experto en evaluación metodológica. La variación de perfiles buscó reducir sesgos al forzar perspectivas independientes sobre el mismo registro.

Cada evaluador respondió de manera binaria si el registro cumplía simultáneamente los

tres criterios de inclusión:

- **C1**: El artículo trata sobre *fine-tuning* de *Large Language Models* (GPT, LLaMA, Mistral, Claude, Falcon, BLOOM, OPT). Se excluyen modelos *encoder-only* como BERT o RoBERTa.
- **C2**: El artículo evalúa al menos una de las nueve técnicas objetivo: *Full Fine-Tuning*, SFT, LoRA, QLoRA, *Adapter Layers*, *Prefix Tuning*, *Prompt Tuning*, RLHF o DPO.
- **C3**: El artículo reporta evaluación empírica con métricas cuantitativas (*accuracy*, F1, BLEU, ROUGE, *perplexity* o *benchmark scores*).

La decisión final se determinó por mayoría simple: con tres o más votos favorables el registro fue aceptado, con dos o menos fue rechazado. Los registros con empate exacto fueron marcados para revisión manual. De los 140 registros evaluados, 101 superaron esta fase y avanzaron a la evaluación de texto completo. Los 39 registros restantes fueron excluidos por no cumplir los criterios de inclusión.

6.2.2 FASE 2: EVALUACIÓN DE TEXTO COMPLETO

Los 101 artículos fueron evaluados en texto completo mediante Claude Sonnet 4.6, utilizando *tool use* estructurado para garantizar respuestas en formato JSON validado. El texto completo de cada PDF fue extraído automáticamente mediante PyMuPDF y enviado al modelo junto con cinco criterios de elegibilidad:

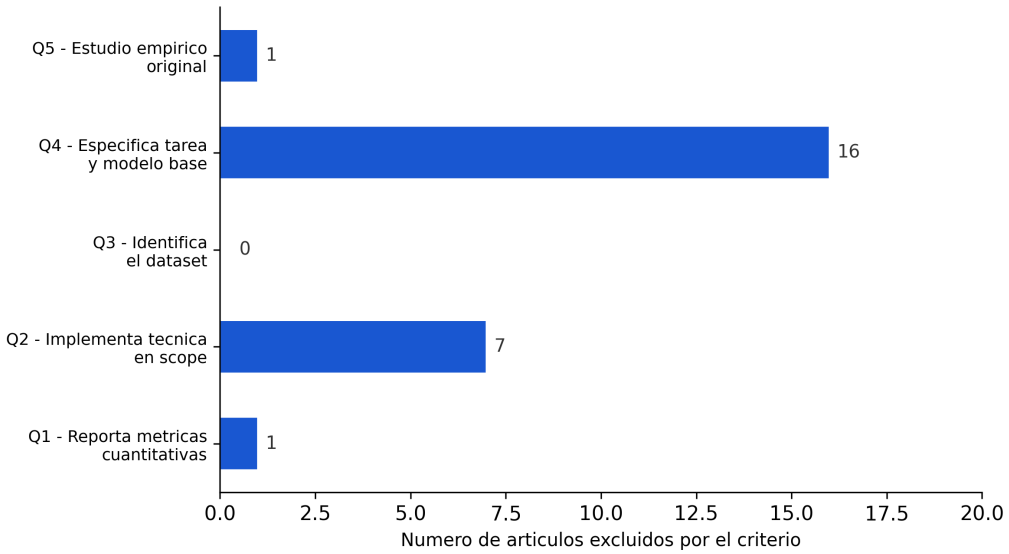
- **Q1**: ¿El artículo reporta resultados numéricos explícitos usando *accuracy*, F1, *precision*, *recall*, BLEU, ROUGE u otros *scores* de *benchmarks* estándar?
- **Q2**: ¿El artículo implementa al menos una técnica en scope con evidencia de aplicación real, reportando *optimizer*, *dataset* y *training steps* o *epochs*?
- **Q3**: ¿El artículo identifica el *dataset* utilizado especificando nombre y dominio de aplicación?

- **Q4**: ¿El artículo especifica tanto la tarea NLP realizada como la arquitectura del modelo base utilizado?
- **Q5**: ¿Es un estudio empírico original con resultados experimentales propios, excluyendo *surveys*, revisiones de literatura y artículos teóricos?

Los criterios Q1, Q2, Q4 y Q5 operaron con lógica eliminatoria directa: un resultado negativo en cualquiera resultó en exclusión inmediata. El criterio Q3 recibió tratamiento especial: un resultado negativo derivó en revisión manual en lugar de exclusión directa, dado que estudios de dominio aplicado pueden describir correctamente su corpus sin nombrar un *dataset* público estándar. De los 101 artículos evaluados, 84 fueron incluidos y 17 excluidos. La Figura 6.2 desglosa las exclusiones por criterio, evidenciando que Q4 fue el más restrictivo con 16 exclusiones, seguido de Q2 con 7. Los criterios Q1 y Q5 generaron una exclusión cada uno, mientras que Q3 no produjo exclusiones directas dado que los casos negativos derivaron en revisión manual.

Figura 6.2

Artículos excluidos por cada criterio de elegibilidad en la Fase 2 (n = 101)



Nota. El criterio Q3 no aparece con exclusiones directas porque los casos negativos derivaron en revisión manual. Los valores pueden sumar más de 17 si un mismo artículo no cumplió más de un criterio.

6.3 EVALUACIÓN DE CALIDAD METODOLÓGICA

Los 84 estudios incluidos fueron sometidos a una evaluación de calidad metodológica mediante Claude Haiku 4.5, utilizando *tool use* estructurado sobre el texto completo de cada artículo. Cada estudio fue evaluado según tres criterios en escala 0 / 0.5 / 1, con un score máximo de 3,0:

- **QA1**: ¿El artículo reporta explícitamente las proporciones del *split* train/test o los tamaños exactos de cada subconjunto?
- **QA3**: ¿El artículo reporta una configuración de entrenamiento completa incluyendo *learning rate*, *batch size*, *epochs* y parámetros técnica-específicos?
- **QA4**: ¿El artículo caracteriza completamente el *dataset* con nombre, dominio y tamaño?

Los scores obtenidos en esta evaluación se emplean como criterio de priorización interpretativa en la síntesis comparativa presentada en el capítulo de Resultados: los hallazgos provenientes de estudios con score alto (score $\geq$ 2,5) se priorizaron como soporte principal de los patrones transversales identificados, mientras que los hallazgos de estudios con calidad media se reportan en el análisis descriptivo pero no fundamentan las conclusiones comparativas.

6.4 CLASIFICACIÓN POR DOMINIO DE APLICACIÓN

Cada uno de los 84 estudios incluidos fue asignado a uno de cuatro dominios de aplicación definidos a priori. La clasificación se realizó de forma automatizada durante la Fase 3, como parte del mismo proceso de evaluación de calidad metodológica, utilizando el texto completo de cada artículo como insumo. Los criterios de asignación empleados fueron los siguientes:

- **Salud y Biomedicina**: estudios cuya tarea principal involucra NLP clínico, registros médicos, texto biomédico, descubrimiento de fármacos, salud mental, radiología, genómica o datos de pacientes.

- **Ciencias Exactas y Computación**: estudios enfocados en generación de código, razonamiento matemático, ingeniería de software, verificación formal, tareas de programación, ciberseguridad o *benchmarks* de ciencias de la computación.
- **Lenguaje y Humanidades**: estudios que abordan tareas generales de NLP como traducción, resumen, análisis de sentimiento, reconocimiento de entidades nombradas, respuesta a preguntas sobre corpus generales, lingüística, literatura o texto de redes sociales.
- **Industria y Negocios**: estudios con aplicaciones en finanzas, derecho, educación, atención al cliente, comercio electrónico, recursos humanos, logística o aplicaciones empresariales.

Cuando un estudio evaluaba múltiples tareas en dominios distintos, se asignó al dominio principal según la contribución declarada por los autores. Esta clasificación constituye el eje organizador del análisis comparativo presentado en el capítulo de Resultados.

6.5 EXTRACCIÓN DE DATOS

De cada uno de los 84 estudios incluidos en el corpus final se extrajeron datos estructurados mediante lectura directa del texto completo de cada artículo. Este proceso se realizó de forma independiente al pipeline de selección PRISMA descrito en las secciones anteriores. Como apoyo a la identificación inicial, se empleó un modelo de lenguaje (Claude, Anthropic) para localizar en cada artículo la técnica principal implementada, la tarea de aplicación y la presencia o ausencia de un *baseline* de referencia explícito; los valores finales fueron verificados y registrados manualmente contrastando cada campo contra el texto original del artículo.

Los campos extraídos fueron los siguientes: técnica de *fine-tuning* implementada, tarea y dominio de aplicación, métrica principal de evaluación, valor del *baseline* de referencia y mejor resultado obtenido por el modelo evaluado. A partir de los dos últimos campos se calculó el incremento porcentual ($\Delta\,\%$) respecto al *baseline*. Los campos no reportados explícitamente en el artículo fueron registrados como *not reported*, sin inferencia ni estimación de valores ausentes.

Para que el cálculo de mejora relativa fuera metodológicamente válido, se aplicó un criterio de comparabilidad: un estudio se consideró comparable únicamente si los autores declaraban

un *baseline* único y explícito contra el cual comparar el resultado principal. Quedaron excluidos de la comparación cuantitativa los estudios que evaluaban múltiples modelos sin designar una referencia de comparación fija, y aquellos donde se presentaban resultados para más de un modelo *fine-tuneado* sin indicar cuál constituía el resultado principal, impidiendo el cálculo de un incremento atribuible a la técnica de forma aislada. Bajo este criterio, 30 de los 84 estudios presentaron pares de valores comparables; los 54 restantes se reportan en el análisis descriptivo del corpus pero no contribuyen a la síntesis cuantitativa de mejora relativa.

6.6 ANÁLISIS DE DATOS

El análisis de datos se organizó en tres niveles complementarios sobre el corpus de 84 estudios incluidos.

El análisis descriptivo caracterizó el corpus mediante frecuencias de técnicas de *fine-tuning*, distribución por dominio de aplicación, modelos base más utilizados y métricas de evaluación más frecuentes. Este nivel permitió identificar los patrones dominantes en la literatura reciente sobre *fine-tuning* de LLMs.

El análisis comparativo sintetizó el rendimiento reportado por dominio de aplicación, comparando los resultados de cada estudio contra sus *baselines* de referencia dentro de cada dominio. Dado que los estudios emplean métricas y *datasets* heterogéneos, la comparación se realizó a nivel de mejora relativa sobre el *baseline* en lugar de valores absolutos, permitiendo identificar patrones de efectividad por dominio independientes de la escala de cada métrica.

El análisis de calidad metodológica empleó los *scores* obtenidos en la evaluación de calidad como criterio de priorización interpretativa: los hallazgos provenientes de estudios con *score* alto ($score \geq 2{,}5$) se priorizaron como soporte principal de los patrones transversales identificados, mientras que los hallazgos de estudios con calidad media se reportan en el análisis descriptivo pero no fundamentan las conclusiones comparativas.

7. RESULTADOS Y DISCUSIÓN

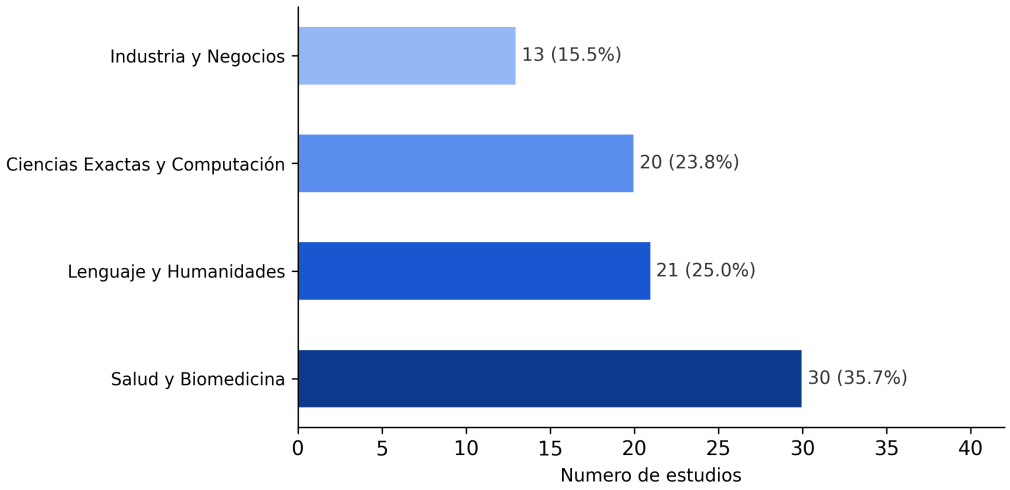
7.1 CARACTERIZACIÓN DEL CORPUS

El proceso de selección PRISMA resultó en la inclusión de 84 estudios a partir de 140 registros identificados en Scopus. Las exclusiones en la Fase 2 correspondieron principalmente a artículos que no especificaban la tarea NLP ni el modelo base (Q4, n = 16), con casos como predicción de trayectorias, clasificación de señales de sensores y regresión numérica sobre datos tabulares, seguido de técnicas fuera del alcance del estudio (Q2, n = 7) y ausencia de resultados experimentales propios (Q5, n = 1).

La clasificación de los 84 estudios en cuatro dominios de aplicación se realizó de forma automatizada durante la Fase 3, aplicando los criterios definidos en la sección de Metodología. La distribución resultante evidencia un marcado predominio del área de Salud y Biomedicina, con 30 estudios (35,7 %), seguido de Lenguaje y Humanidades con 21 (25,0 %), Ciencias Exactas y Computación con 20 (23,8 %) e Industria y Negocios con 13 (15,5 %). Esta distribución refleja la intensa actividad investigativa que el *fine-tuning* de *Large Language Models* ha tenido en aplicaciones clínicas y biomédicas durante el período 2020–2025, fenómeno concordante con la disponibilidad de corpus especializados como PubMed y MIMIC y con la demanda de modelos adaptados para tareas de extracción de información médica. La Figura 7.1 presenta esta distribución de forma visual.

Figura 7.1

Distribución del corpus por dominio de aplicación (n = 84)



La calidad metodológica del corpus fue aceptable en conjunto: el *score* promedio fue de 2,61 sobre 3,0, con la mayoría de estudios por encima del umbral de priorización de 2,5 establecido en la Metodología. A nivel de criterios individuales, QA4 obtuvo la puntuación media más alta, reflejando que prácticamente la totalidad de los estudios caracteriza con precisión sus *datasets*. QA1 y QA3 presentaron medias más bajas, con una proporción menor de estudios que reportan *splits* explícitos o que detallan de forma completa los hiperparámetros de entrenamiento. Por dominio, Lenguaje y Humanidades concentra la mayor calidad metodológica promedio, mientras que Industria y Negocios presenta la media más baja con la mayor variabilidad, lo que refleja heterogeneidad en las prácticas de reporte dentro de ese dominio.

7.2 RESULTADOS POR DOMINIO

7.2.1 SALUD Y BIOMEDICINA

El dominio de Salud y Biomedicina concentra 30 estudios, constituyendo el grupo más numeroso del corpus (35,7 %).

TÉCNICAS Y MODELOS PREDOMINANTES

En este dominio, LoRA y QLoRA constituyen la categoría dominante con 21 de los 30 estudios (70,0 %), seguidas de técnicas alternativas como Adapters, Prefix Tuning y variantes con 5 estudios (16,7 %), *Full Fine-Tuning* con 3 (10,0 %) y métodos de alineación con 1 (3,3 %). Dentro de la familia LoRA, dos estudios incorporan GRPO como mecanismo de alineación: educación cardiovascular (Rullo et al., 2025) y generación de reportes clínicos (Kuo et al., 2025). El único método de alineación es SFT+DPO para razonamiento clínico (Savage et al., 2025). Técnicas alternativas como P-Tuning v2 (C. Liu et al., 2024) y *Prompt Tuning* (W. Yang et al., 2025) aparecen de forma puntual, reflejando que el dominio biomédico experimenta con distintos enfoques de adaptación paramétrica aunque LoRA concentra la mayoría de las aplicaciones.

TAREAS Y MÉTRICAS

Las tareas identificadas en el dominio se agrupan en cuatro categorías principales. La extracción de información clínica (reconocimiento de entidades nombradas, extracción de medicamentos y relaciones) es la más frecuente, con estudios en NER médico en chino (L. Zhou et al., 2024) y extracción de medicamentos (Richter-Pechanski et al., 2025). Le siguen la respuesta a preguntas médicas y los sistemas de diálogo especializado (Adhikary et al., 2025; Bayram et al., 2025; Tong et al., 2025), la generación y resumen de texto clínico (Chao et al., 2025; Hou et al., 2024) y la clasificación en salud pública, que abarca análisis de sentimiento (Elmitwalli et al., 2024) y detección de desinformación (Zhao et al., 2025).

La heterogeneidad de métricas reportadas refleja directamente la diversidad de subtareas. En tareas de extracción y clasificación predominan F1 y sus variantes; en NER médico, por ejemplo, se reporta una mejora de F1 de 62,40 % a 65,90 % (L. Zhou et al., 2024). En tareas de generación y resumen predominan ROUGE y BLEU, con ejemplos como la generación de cartas médicas (Hou et al., 2024). Una limitación recurrente en el corpus es la ausencia de un *baseline* declarado explícitamente: varios estudios evalúan múltiples modelos o configuraciones y reportan el mejor resultado sin establecer una referencia de comparación fija, lo que impide atribuir la mejora a la técnica de *fine-tuning* de forma aislada. De los 30 estudios, 9 reportan métricas cuantitativas comparables; los resultados detallados por estudio se presentan en la Tabla 10.1

del Apéndice A.

PATRONES DE APLICABILIDAD

Del análisis de los 30 estudios emergen tres patrones de aplicabilidad consistentes en el dominio.

El primero es la ventaja del *fine-tuning* específico sobre el modelo base sin adaptar. Los estudios con resultados numéricos explícitos muestran que LoRA y QLoRA producen mejoras sustanciales sobre los *baselines* de referencia en las subtareas evaluadas. Este patrón se observa en el dominio geriátrico, donde LLaMA3.1-8B ajustado sobre datos de cuidado especializado mejora de ROUGE 8,60 % a 34,96 % (+306,5 %) respecto al mismo modelo sin adaptar (J. Xu et al., 2025), y en generación de cartas médicas, donde QLoRA mejora ROUGE-1 de 0,161 a 0,352 (+118,6 %) (Hou et al., 2024).

El segundo es la preferencia por modelos abiertos en entornos regulados. La imposibilidad de enviar datos clínicos sensibles a APIs externas condiciona directamente la elección de técnica y arquitectura. Al menos un estudio señala explícitamente que los modelos propietarios no pudieron ser evaluados porque las versiones seguras para datos clínicos protegidos no estaban disponibles al momento del estudio (Chao et al., 2025). Este condicionante favorece el despliegue local mediante técnicas como QLoRA o LoRA estándar, que no requieren dependencia de servicios externos.

El tercero es la efectividad de LoRA en lenguas con escasa representación en los corpus de preentrenamiento. En idiomas como turco, vietnamita y árabe médico, donde los modelos base tienen menor exposición previa, LoRA produce adaptaciones consistentes al dominio clínico (Bayram et al., 2025; Bui et al., 2025; Khader et al., 2025). El mismo comportamiento se observa en dominios de conocimiento altamente especializado como la medicina tradicional china (Tong et al., 2025; G. Yang et al., 2024).

7.2.2 LENGUAJE Y HUMANIDADES

El dominio de Lenguaje y Humanidades concentra 21 estudios (25,0 %) y presenta la mayor calidad metodológica promedio entre los cuatro dominios.

TÉCNICAS Y MODELOS PREDOMINANTES

LoRA y QLoRA constituyen también la categoría dominante con 16 de los 21 estudios (76,2 %), seguidas de técnicas alternativas como *Adapter Layers* (K. Chen et al., 2025), Multi-PEFT (Jia et al., 2025) y *Prompt Tuning* (Cheng y Cao, 2025) con 3 estudios (14,3 %), y métodos de alineación con 2 (9,5 %): DPO para Q&A en kazajo (Kadyrbek et al., 2025) y ORPO para análisis de sentimiento (J. Shen et al., 2025). Esta distribución refleja mayor diversidad metodológica que otros dominios del corpus.

TAREAS Y MÉTRICAS

Las tareas en este dominio son heterogéneas y reflejan la amplitud del área. Las más frecuentes son análisis de sentimiento y detección de lenguaje ofensivo (Rathore y Chaurasia, 2025; J. Shen et al., 2025; Y. Wang et al., 2025; Zain et al., 2025), seguidas de respuesta a preguntas y razonamiento (Z. Chen y Liu, 2025; Cheng y Cao, 2025; H. Li et al., 2025). Otras tareas identificadas son traducción automática y estimación de calidad (Khoboko et al., 2025; Sindhujan et al., 2025) y evaluación educativa automatizada (Anghel et al., 2025).

Las métricas varían según la subtarea. F1 y *accuracy* predominan en clasificación; en lenguaje ofensivo, por ejemplo, se reporta una mejora de Macro-F1 de 80,1 a 81,3 (+1,5 %) (Y. Wang et al., 2025). En respuesta a preguntas predominan métricas de solapamiento como ROUGE-L, con una mejora de 22,39 % a 47,85 % (+113,7 %) en Q&A sobre patrimonio cultural (Z. Chen y Liu, 2025). De los 21 estudios, 9 reportan métricas cuantitativas comparables; los resultados detallados por estudio se presentan en la Tabla 10.2 del Apéndice A.

PATRONES DE APLICABILIDAD

Del análisis de los 21 estudios emergen dos patrones de aplicabilidad predominantes en el dominio.

El primero es la efectividad de las técnicas de *fine-tuning* eficiente en lenguas de bajos recursos y dominios culturalmente específicos. Varios estudios demuestran que LoRA, QLoRA y métodos de alineación como DPO permiten adaptar modelos generales a lenguas con escasa representación en los datos de preentrenamiento, como kazajo y zulú, con mejoras consistentes sobre los *baselines* (Kadyrbek et al., 2025; Khoboko et al., 2025). Este patrón refuerza la aplicabilidad transversal de las técnicas PEFT en escenarios de recursos limitados y complementa lo observado en Salud y Biomedicina con lenguas como turco, vietnamita y árabe.

El segundo es la diversificación de enfoques PEFT orientados a tareas específicas. La aparición de Multi-PEFT, *Adapter Layers*, *Prompt Tuning* y métodos de alineación como ORPO en este dominio sugiere que la investigación en Lenguaje y Humanidades está explorando combinaciones y variantes más allá de LoRA estándar. Arquitecturas agnósticas a la tarea que combinan múltiples estrategias PEFT superan a LoRA estándar en *benchmarks* de clasificación y razonamiento, con una mejora de Macro-F1 de 0,8712 a 0,8736 sobre el mejor *baseline* disponible (Jia et al., 2025), y Adapter Layers mejoran el rendimiento promedio en clasificación GLUE de 86,4 a 89,0 (+3,0 %) (K. Chen et al., 2025).

7.2.3 CIENCIAS EXACTAS Y COMPUTACIÓN

El dominio de Ciencias Exactas y Computación concentra 20 estudios (23,8 %).

TÉCNICAS Y MODELOS PREDOMINANTES

LoRA y QLoRA alcanzan en este dominio su mayor concentración relativa del corpus, con 17 de los 20 estudios (85,0 %), incluyendo combinaciones como LoRA+RLHF (Gorbatovski et al., 2024) y DP-LoRA (X. Liu et al., 2025). Las técnicas alternativas (Prefix Tuning, *Prompt Tuning* y combinaciones múltiples) representan 3 estudios (15,0 %) (Jung et al., 2025; Prottasha

et al., 2024; Zou et al., 2025).

TAREAS Y MÉTRICAS

Las tareas se agrupan en tres categorías principales. La más frecuente es el razonamiento y la respuesta a preguntas sobre *benchmarks* de propósito general, presente en estudios que evalúan técnicas sobre conjuntos como SuperGLUE (J. Liu et al., 2025), GLUE (Jung et al., 2025) y razonamiento de sentido común (Q. Wang y Shen, 2025). Le sigue la ingeniería de software, con estudios que abarcan resumen de reportes de bugs (Choi y Yang, 2025; Xiang y Shao, 2024), generación de código (Uluırmak y Kurban, 2025) y razonamiento sobre grafos (Y. Chen et al., 2025). El aprendizaje federado con técnicas PEFT aparece en cuatro estudios (Elbakary et al., 2025; Iftikhar et al., 2025; L. Li et al., 2025; X. Liu et al., 2025), reflejando un interés creciente en la adaptación de modelos bajo restricciones de privacidad distribuida.

Las métricas más frecuentes son *accuracy* y sus variantes en tareas de razonamiento (Iftikhar et al., 2025; J. Liu et al., 2025). ROUGE predomina en tareas de generación y resumen; en resumen de reportes de bugs, por ejemplo, se reportan mejoras de ROUGE-1 Recall de 61,0 % a 84,6 % (+38,7 %) (Choi y Yang, 2025) y de ROUGE-L de 39,69 a 41,01 (+3,3 %) (Xiang y Shao, 2024). De los 20 estudios, 6 reportan *baseline* explícito que permite calcular mejora relativa, mientras que 14 no incluyen referencia de comparación directa; los resultados detallados por estudio se presentan en la Tabla 10.3 del Apéndice A.

PATRONES DE APLICABILIDAD

Del análisis de los 20 estudios emergen dos patrones de aplicabilidad en el dominio.

El primero es la adopción transversal de LoRA y QLoRA en tareas computacionales de distinta naturaleza. La concentración en 17 de los 20 estudios refleja que esta familia de técnicas se aplica de forma consistente tanto en tareas de razonamiento sobre *benchmarks* como en ingeniería de software. Entre los estudios con métricas comparables, las mejoras van desde incrementos moderados en resumen de reportes de bugs (Xiang y Shao, 2024) hasta ganancias pronunciadas en entrenamiento multi-parte, donde LoRA mejora *accuracy* de 14,25 a 65,43

(+359,2 %) (N. Liu y Liu, 2025), lo que sugiere que la técnica se adapta a contextos experimentales diversos dentro del dominio.

El segundo es la aplicación de técnicas PEFT en escenarios de aprendizaje federado como dirección de investigación emergente. Cuatro estudios exploran la combinación de LoRA con marcos de aprendizaje federado para entrenar modelos sobre datos distribuidos sin centralización (Elbakary et al., 2025; Iftikhar et al., 2025; L. Li et al., 2025; X. Liu et al., 2025). Los resultados disponibles son mixtos: un estudio reporta una mejora de *accuracy* de 71,2 % a 73,8 % (Iftikhar et al., 2025) mientras que otro muestra una leve caída de 76,6 % a 75,4 % (L. Li et al., 2025), lo que indica que esta combinación está aún en fase exploratoria sin evidencia consolidada de efectividad.

7.2.4 INDUSTRIA Y NEGOCIOS

El dominio de Industria y Negocios concentra 13 estudios (15,5 %), siendo el grupo menos numeroso del corpus. La calidad metodológica promedio es la más baja de los cuatro dominios y con la mayor variabilidad, lo que refleja heterogeneidad en las prácticas de reporte dentro del dominio.

TÉCNICAS Y MODELOS PREDOMINANTES

LoRA o alguna de sus variantes aparece en los 13 estudios del dominio (100,0 %), confirmando su posición como categoría predominante. Varios estudios combinan LoRA con técnicas complementarias: con RLHF para generación de patentes (Ren et al., 2025) y con RAG para Q&A sobre protección fitosanitaria (Xiong et al., 2025). La variante MoELoRA se aplica a mantenimiento minero (Cao et al., 2025), y LoRA combinada con RLHF-PPO a generación de texto legal (Peng y Lei, 2025).

TAREAS Y MÉTRICAS

Las tareas son aplicadas y de nicho específico. Las más frecuentes son la generación de texto especializado en dominios verticales, con estudios que abarcan generación de patentes (Ren

et al., 2025), análisis de riesgo en mantenimiento minero (Cao et al., 2025; Sun et al., 2024), consultoría en acuicultura (Song et al., 2025) y generación de texto legal (Peng y Lei, 2025). Le siguen sistemas de Q&A y recuperación de información en dominios industriales (Xiong et al., 2025; W. Zhou et al., 2025); a estos se suman análisis de sentimiento en redes sociales (D. Lin et al., 2024) y extracción geoespacial (Y. Han et al., 2025).

Las métricas reportadas son heterogéneas: ROUGE-L en consultoría de dominio (Song et al., 2025) y Macro-F1 en extracción de información (Y. Yang et al., 2025). De los 13 estudios, 6 reportan *baseline* explícito que permite calcular mejora relativa; los resultados detallados por estudio se presentan en la Tabla 10.4 del Apéndice A.

PATRONES DE APLICABILIDAD

Del análisis de los 13 estudios emergen dos patrones de aplicabilidad característicos del dominio.

El primero es la combinación sistemática de LoRA con técnicas complementarias para tareas de generación de alto valor. En los estudios con resultados cuantitativos más destacados, LoRA no actúa sola sino combinada con RLHF como mecanismo de alineación o con RAG como fuente de conocimiento actualizado. La combinación LoRA+RLHF produce una mejora de BLEU-4 de 11,663 a 45,360 (+288,9 %) en escala 0-100 en generación de patentes (Ren et al., 2025), y LoRA+RAG mejora ROUGE-2 de 0,121 a 0,321 (+165,3 %) en Q&A sobre protección fitosanitaria (Xiong et al., 2025). Este patrón sugiere que en aplicaciones industriales donde la calidad del texto generado es crítica, la combinación de LoRA con mecanismos complementarios es la configuración predominante.

El segundo es la aplicabilidad de LoRA en dominios de conocimiento altamente especializado. En acuicultura (Song et al., 2025), mantenimiento minero (Cao et al., 2025) y gestión de emergencias (W. Zhou et al., 2025), LoRA produce mejoras consistentes en dominios verticales con vocabulario técnico específico y escasa representación en los corpus de preentrenamiento. En NER de emergencias, por ejemplo, se reporta una mejora de F1 de 0,7874 a 0,8920 (+13,3 %) (W. Zhou et al., 2025).

7.3 ANÁLISIS COMPARATIVO ENTRE DOMINIOS

El análisis transversal de los cuatro dominios permite identificar patrones globales en el uso de técnicas de *fine-tuning*, la cobertura metodológica del corpus y las diferencias en rendimiento reportado entre contextos de aplicación.

7.3.1 DISTRIBUCIÓN DE TÉCNICAS

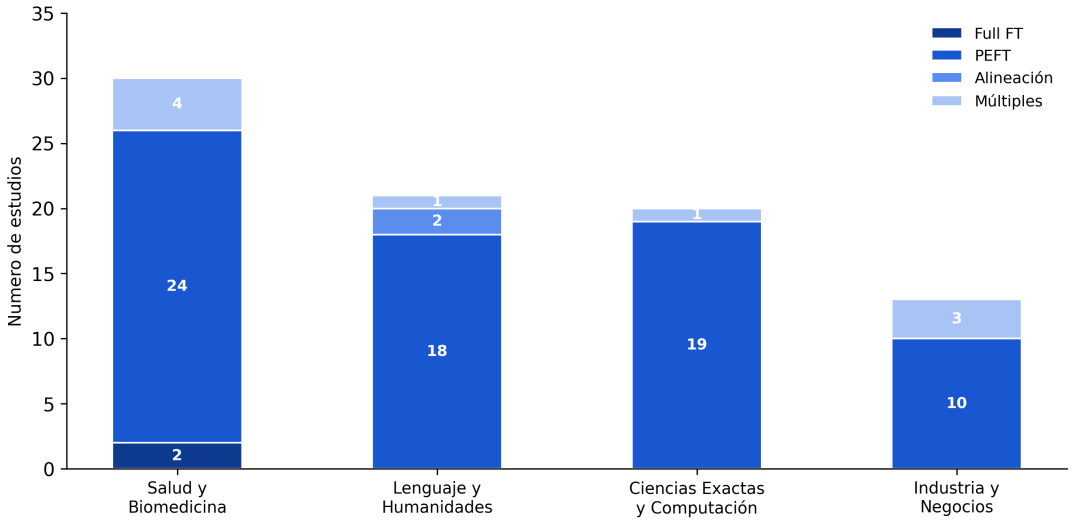
A nivel global, cualquier variante de LoRA o QLoRA aparece en 67 de los 84 estudios del corpus (79,8 %), consolidando esta categoría como la más adoptada independientemente del dominio. Le siguen otras técnicas PEFT sin LoRA ni QLoRA con 11 estudios (13,1 %), *Full Fine-Tuning* e *Instruction Tuning* con 3 (3,6 %) y métodos de alineación sin LoRA con 3 (3,6 %).

La familia PEFT emerge como dominante en todos los dominios del corpus, con LoRA como técnica principal. La concentración de LoRA/QLoRA crece progresivamente desde Salud y Biomedicina (70,0 %) hacia Industria y Negocios (100,0 %), lo que sugiere que cuanto más especializado y orientado a producción es el contexto, mayor es la convergencia hacia LoRA como técnica estándar. Los métodos de alineación mantienen una presencia marginal pero consistente en dos de los cuatro dominios, mientras que *Full Fine-Tuning* y SFT quedan restringidos a los dominios con mayor demanda de adaptación completa del modelo.

La Figura 7.2 presenta la distribución por familia de técnicas en cada dominio, y la Figura 7.3 desglosa las técnicas específicas identificadas en el corpus.

Figura 7.2

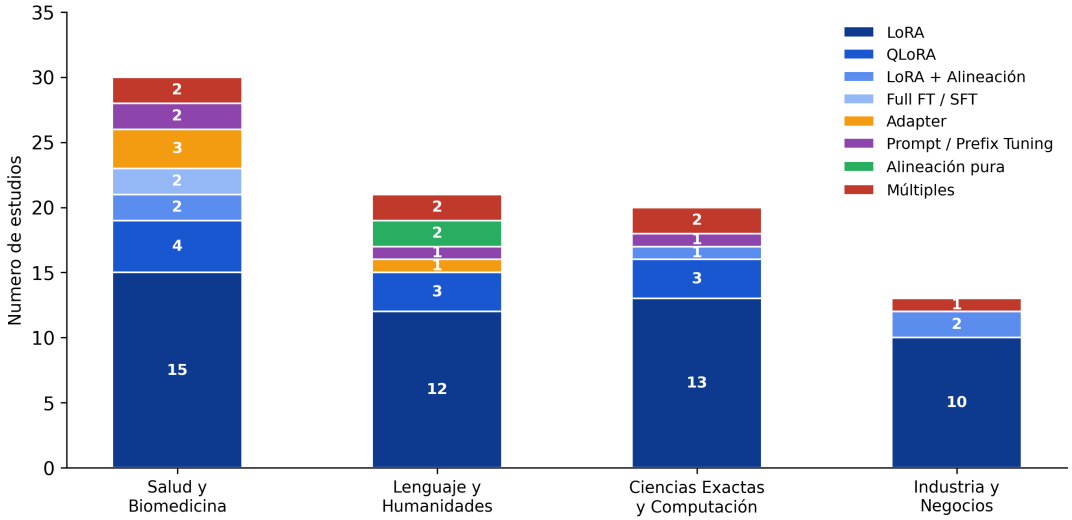
Distribución por familia de técnicas de fine-tuning por dominio (n = 84)



Nota. Las familias corresponden a las tres categorías definidas en el marco teórico: *Full Fine-Tuning*, PEFT y Alineación. Un estudio puede contribuir a más de una familia si combina técnicas de distintas categorías.

Figura 7.3

Uso de técnicas específicas de fine-tuning por dominio de aplicación (n = 84)



Nota. "LoRA + Alineación" agrupa estudios que combinan LoRA con RLHF, DPO u ORPO como mecanismo complementario. "Alineación pura" corresponde a estudios que aplican DPO u ORPO sin LoRA como técnica base. "Múltiples" agrupa estudios que combinan técnicas de distintas familias sin una técnica base dominante. Un estudio puede contabilizarse en más de una categoría si combina técnicas.

La forma en que LoRA se combina con otras técnicas varía según el dominio. En Salud y Biomedicina aparece combinada con GRPO en reportes clínicos (Kuo et al., 2025) y con otros mecanismos de adaptación complementaria. En Industria y Negocios predomina su combinación con RLHF para alineación (Ren et al., 2025) y con RAG para recuperación de conocimiento especializado (Xiong et al., 2025). En contraste, Lenguaje y Humanidades y Ciencias Exactas y Computación presentan mayor proporción de estudios que usan LoRA de forma aislada, reflejando un carácter más experimental y comparativo (Anghel et al., 2025; Z. Chen y Liu, 2025; Xiang y Shao, 2024).

7.3.2 COBERTURA METODOLÓGICA Y REPORTE DE MÉTRICAS

La proporción de estudios con pares de valores comparables (baseline y resultado explícitos) varía entre dominios. Salud y Biomedicina y Lenguaje y Humanidades presentan 9 estudios comparables cada uno (30,0 % y 42,9 % de sus respectivos dominios), Industria y Negocios 6 de 13 (46,2 %) y Ciencias Exactas y Computación 6 de 20 (30,0 %). La ausencia de valores comparables es más frecuente en dominios donde las contribuciones se centran en el diseño de sistemas o la evaluación cualitativa, como ocurre en Salud y en CyC, donde los jueces expertos o la ausencia de un único *baseline* de referencia impiden el cálculo de mejora relativa.

Las métricas empleadas difieren según el dominio. F1 predomina en Salud y Biomedicina dado el énfasis en extracción de entidades y clasificación clínica (L. Zhou et al., 2024). *Accuracy* es la métrica más frecuente en Ciencias Exactas y Computación, donde los *benchmarks* de razonamiento dominan (J. Liu et al., 2025). En Lenguaje y Humanidades coexisten F1 (S. Han et al., 2025), *accuracy* (Kadyrbek et al., 2025; J. Shen et al., 2025) y métricas específicas de tarea como ROUGE-L en generación (Z. Chen y Liu, 2025). En Industria y Negocios la heterogeneidad es máxima, con BLEU-4 (Ren et al., 2025), Macro-F1 (Y. Yang et al., 2025) y ROUGE-L (Song et al., 2025) distribuidos según la naturaleza de cada aplicación vertical.

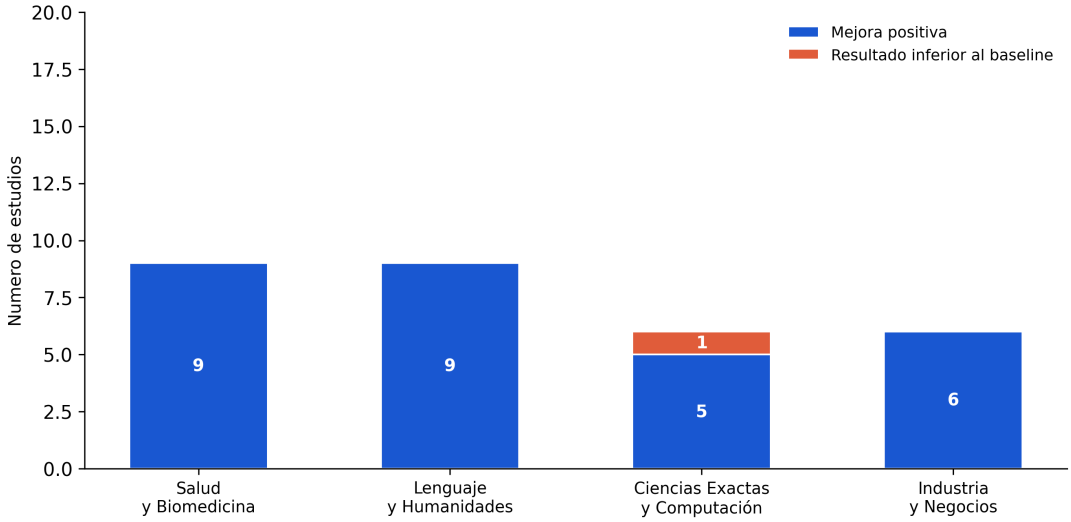
7.3.3 PATRONES TRANSVERSALES

Tres patrones emergen de forma consistente a través de los cuatro dominios del corpus. La Figura 7.4 resume la dirección de los resultados reportados, mostrando que la gran mayoría de

los estudios con métricas comparables reportan mejora positiva sobre sus *baselines* de referencia.

Figura 7.4

Dirección de la mejora respecto al baseline por dominio (estudios con pares de valores comparables, n = 30)



Nota. Se incluyen únicamente los 30 estudios del corpus que declaran explícitamente un *baseline* de referencia y reportan tanto el valor de referencia como el resultado obtenido tras el *fine-tuning*, lo que permite calcular la dirección y magnitud del cambio. El único resultado negativo corresponde al dominio de Ciencias Exactas y Computación.

El primer patrón es la dominancia transversal de LoRA como técnica de adaptación. LoRA o alguna de sus variantes está presente en 67 de los 84 estudios del corpus (79,8 %), con una frecuencia que se mantiene alta en todos los dominios sin excepción. Este patrón de adopción no depende de la tarea, el idioma ni el tamaño del modelo, lo que sugiere que LoRA ha consolidado su posición como técnica de referencia en la literatura de *fine-tuning* de LLMs durante el período analizado.

El segundo patrón es la efectividad consistente del *fine-tuning* sobre el modelo base sin adaptar. De los 30 estudios con pares de valores comparables, la dirección del cambio es positiva en 29 casos, con mejoras relativas que en varios estudios superan el 100 %. Este patrón se observa en dominios y tareas diversas: en diálogo geriátrico se reporta una mejora de ROUGE de 8,60 % a 34,96 % (+306,5 %) (J. Xu et al., 2025), y en extracción de información militar una ganancia

de Macro-F1 de 0,2140 a 0,7020 (+228,0 %) (Y. Yang et al., 2025). Es importante señalar que la heterogeneidad de métricas y *datasets* entre estudios impide una agregación directa de valores absolutos; el patrón se establece a nivel de dirección de la mejora, no de magnitud promedio entre dominios.

El tercer patrón es la tendencia hacia arquitecturas compuestas en aplicaciones orientadas a producción. La familia PEFT, con LoRA como técnica predominante, no solo domina en frecuencia sino que en contextos aplicados tiende a combinarse con mecanismos complementarios: con RLHF para alineación de preferencias (Ren et al., 2025), con RAG para recuperación de conocimiento especializado (Xiong et al., 2025) y con marcos de aprendizaje federado para privacidad distribuida (Elbakary et al., 2025; X. Liu et al., 2025). Este patrón sugiere que la familia PEFT actúa como componente base sobre el que se construyen sistemas más complejos en aplicaciones de producción, especialmente en los dominios de Salud y Biomedicina e Industria y Negocios.

8. CONCLUSIONES

La presente revisión sistemática analizó 84 estudios empíricos publicados entre 2020 y 2025 sobre técnicas de *fine-tuning* para *Large Language Models*, seleccionados mediante el protocolo PRISMA desde una búsqueda inicial de 140 registros en Scopus. Los hallazgos se organizan en correspondencia con los tres objetivos específicos planteados.

En relación con el primer objetivo, sistematizar la literatura mediante PRISMA extrayendo características técnicas, métricas y resultados por dominio, el corpus presenta una distribución heterogénea entre cuatro dominios de aplicación: Salud y Biomedicina (30 estudios), Lenguaje y Humanidades (21), Ciencias Exactas y Computación (20) e Industria y Negocios (13). LoRA y sus variantes, incluyendo QLoRA, emergen como la categoría dominante con presencia en el 79,8 % de los estudios (67 de 84). La calidad metodológica del corpus fue aceptable en conjunto, con un *score* promedio de 2,61 sobre 3,0. La extracción de datos permitió documentar técnicas, modelos base, *datasets*, métricas y resultados de forma estructurada y trazable para cada uno de los 84 estudios incluidos, cumpliendo los requisitos de reproducibilidad del protocolo PRISMA.

En relación con el segundo objetivo, analizar el rendimiento por dominio, los *trade-offs* y los patrones de aplicabilidad, el análisis revela que el *fine-tuning* paramétrico eficiente produce mejoras consistentes sobre el modelo base sin adaptar en todos los dominios evaluados, aunque con magnitudes variables según el contexto. Los estudios con métricas comparables muestran que las mejoras más consistentes se concentran en tareas de baja cobertura lingüística y dominios altamente especializados, donde el ajuste fino sobre datos del dominio produce ganancias sustanciales independientemente del tamaño del modelo. Los patrones de aplicabilidad difieren entre dominios: en Salud y Biomedicina predomina la preferencia por modelos abiertos motivada por restricciones de privacidad; en Lenguaje y Humanidades destaca la diversificación de variantes de LoRA orientadas a tareas específicas; en Ciencias Exactas y Computación LoRA se consolida como referencia establecida e integrada con aprendizaje federado; y en Industria y Negocios la configuración dominante combina LoRA con RLHF para tareas de generación de alto valor.

En relación con el tercer objetivo, sintetizar criterios de selección según contexto operativo, limitaciones y direcciones futuras, la evidencia acumulada permite formular tres criterios basados en evidencia empírica. Cuando existen restricciones de privacidad o se opera sobre datos sensibles, QLoRA sobre modelos abiertos es la configuración más respaldada por la literatura, con documentación explícita en estudios clínicos y biomédicos. Cuando el dominio de aplicación es altamente especializado o la lengua tiene baja cobertura en los datos de preentrenamiento, LoRA con *datasets* específicos del dominio produce las mejoras más sustanciales independientemente del tamaño del modelo base. Cuando el objetivo es la calidad y alineación del texto generado en aplicaciones de producción, la combinación LoRA + RLHF es la configuración predominante en la literatura reciente, especialmente en dominios industriales y legales.

En respuesta a la pregunta de investigación, las técnicas de *fine-tuning* para *Large Language Models* desarrolladas entre 2020 y 2025 convergen hacia un modelo de adaptación paramétrica eficiente con LoRA como componente central, complementado con mecanismos adicionales según el contexto operativo. La selección de técnica no depende exclusivamente del rendimiento absoluto sino de factores contextuales como la privacidad de los datos, la disponibilidad de recursos lingüísticos del dominio y los requisitos de alineación de las respuestas generadas. Este patrón, consistente a través de dominios y tareas, sugiere que la comunidad investigadora ha convergido hacia un conjunto acotado de configuraciones efectivas durante el período analizado.

La presente revisión presenta cuatro limitaciones principales que deben considerarse al interpretar sus hallazgos. La primera es la restricción a una única base de datos: la búsqueda se realizó exclusivamente en Scopus, lo que puede haber excluido estudios relevantes indexados únicamente en IEEE Xplore, ACM Digital Library o disponibles como preprints en arXiv. La segunda es la heterogeneidad de *baselines*: los estudios incluidos emplean referencias de comparación de naturaleza diversa, lo que impide una síntesis cuantitativa unificada y limita la comparación directa de magnitudes de mejora entre estudios y dominios. La tercera es la ausencia de *baseline* explícito en el 64,3 % de los estudios incluidos (54 de 84), lo que concentra las comparaciones en el subconjunto con reporte completo. La cuarta es el sesgo de acceso abierto: el filtro aplicado para garantizar la reproducibilidad del proceso de extracción puede haber introducido un sesgo de selección hacia ciertos tipos de publicaciones o dominios con mayor tradición de publicación abierta.

Los hallazgos de esta revisión sugieren tres direcciones prioritarias para la investigación futura. La primera es la estandarización de protocolos de evaluación y reporte: la heterogeneidad de métricas, *baselines* y configuraciones experimentales documentada en este corpus dificulta la acumulación de evidencia comparativa, y el desarrollo de estándares análogos a los existentes en otras áreas de *machine learning* permitiría síntesis cuantitativas más robustas. La segunda es la investigación en escenarios de recursos limitados: los hallazgos más consistentes del corpus corresponden a adaptaciones en lenguas de baja cobertura y dominios con datos escasos, y la investigación futura debería profundizar en el tamaño mínimo de *dataset* necesario para obtener mejoras significativas y en la sensibilidad a la calidad de los datos de entrenamiento. La tercera es la evaluación comparativa de técnicas de alineación: RLHF y DPO aparecen en el corpus principalmente en dominios aplicados pero con escasa evidencia cuantitativa comparativa entre ellas, y estudios que las comparen sistemáticamente con métricas y *baselines* compartidos contribuirían a orientar la selección de técnica en aplicaciones de producción.

9. REFERENCIAS

Adhikary, P. K., Motiyani, I., Oke, G., Joshi, M., Pathak, K., Singh, S. M. y Chakraborty, T. (2025). Menstrual Health Education Using a Specialized Large Language Model in India: Development and Evaluation Study of MenstLLaMA. *Journal of Medical Internet Research*, *27*. https://doi.org/10.2196/71977

Anghel, C., Crăciun, M. V., Pecheanu, E., Cocu, A., Anghel, A. A., Iacobescu, P., Maier, C., Andrei, C. A., Scheau, C. y Dragosloveanu, S. (2025). CourseEvalAI: Rubric-Guided Framework for Transparent and Consistent Evaluation of Large Language Models. *Computers*, *14*(10). https://doi.org/10.3390/computers14100431

Ao, S., Dong, Y., Hu, J. y Ramchurn, S. D. (2025). Safe Pruning LoRA: Robust Distance-Guided Pruning for Safety Alignment in Adaptation of LLMs. *Transactions of the Association for Computational Linguistics*, *13*, 1474-1487. https://doi.org/10.1162/TACL.a.44

Bao, X., Zhu, X., Yang, D., Lou, H., Wang, R., Wu, Y., Li, W., Xia, Y., Zeng, L., Pan, Y., Wang, X., Zhang, X., Ling, C., Ling, Y., Zhang, Y., Zhao, Q. y Yang, M. (2025). eCBT-I dialogue system: a comparative evaluation of large language models and adaptation strategies for insomnia treatment. *Journal of Translational Medicine*, *23*(1). https://doi.org/10.1186/s12967-025-06871-y

Bayram, M. A., Diri, B. e Yıldırım, S. (2025). Healthcare-Focused Turkish Medical LLM: Training on Real Patient-Doctor Question-Answer Data for Enhanced Medical Insight. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *24*(11). https://doi.org/10.1145/3772000

Bui, N., Nguyen, G. T., Nguyen, N., Vo, B., Vo, L., Huynh, T., Tang, A. K., Tran, V. N., Huynh, T., Nguyen, H. Q. y Dinh, M. N. (2025). Fine-tuning large language models for improved health communication in low-resource languages. *Computer Methods and Programs in Biomedicine*, *263*. https://doi.org/10.1016/j.cmpb.2025.108655

Cao, X., Wang, X., Shi, L., Yang, X., Zhang, X. y Duan, Y. (2025). A Construction Method for a Coal Mining Equipment Maintenance Large Language Model Based on Multi-Dimensional Prompt Learning and Improved LoRA. *Mathematics*, *13*(10). https://doi.org/10.3390/math13101638

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q. y Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3). https://doi.org/10.1145/3641289

Chao, C. J., Banerjee, I., Arsanjani, R., Ayoub, C., Tseng, A. S., Delbrouck, J. B., Kane, G. C., López-Jiménez, F., Attia, Z. I., Oh, J. K., Erickson, B. J. M., Fei-Fei, L., Adeli-Mosabbeb, E. y Langlotz, C. P. (2025). Evaluating large language models in echocardiography reporting: opportunities and challenges. *European Heart Journal - Digital Health*, *6*(3), 326-339. https://doi.org/10.1093/ehjdh/ztae086

Chen, K., Peng, C., He, X., Sun, J., Liu, X., Qin, X. y Zhong, Y. (2025). AdaptForever: Elastic and Mutual Learning for Continuous NLP Task Mastery. *Computers, Materials and Continua*, *82*(3), 4003-4019. https://doi.org/10.32604/cmc.2025.057443

Chen, Y., Duan, X. y Guo, Y. (2025). GLR: Graph Chain-of-Thought with LoRA Fine-Tuning and Confidence Ranking for Knowledge Graph Completion. *Applied Sciences (Switzerland)*, *15*(13). https://doi.org/10.3390/app15137282

Chen, Z. y Liu, B. (2025). Enhancing Ancient Ceramic Knowledge Services: A Question Answering System Using Fine-Tuned Models and GraphRAG. *Information (Switzerland)*, *16*(9). https://doi.org/10.3390/info16090792

Cheng, Z. y Cao, Y. (2025). Prompt-Based Learning for Factual Knowledge Infusion in Large Language Models. *International Journal of Crowd Science*, *9*(4), 262-268. https://doi.org/10.26599/IJCS.2025.9100014

Chitrakala, S. y Pavithra, S. S. (2025). RL-TweetGen: A Socio-Technical Framework for Engagement-Optimized Short Text Generation in Digital Commerce Using Large Language Models and Reinforcement Learning. *Journal of Theoretical and Applied Electronic Commerce Research*, *20*(3). https://doi.org/10.3390/jtaer20030218

Choi, S. y Yang, G. (2025). AgentReport: A Multi-Agent LLM Approach for Automated and Reproducible Bug Report Generation. *Applied Sciences (Switzerland)*, *15*(22). https://doi.org/10.3390/app152211931

Dettmers, T., Pagnoni, A., Holtzman, A. y Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.2305.14314

Dietrich, J. y Hollstein, A. (2025). Performance and Reproducibility of Large Language Models in Named Entity Recognition: Considerations for the Use in Controlled Environments. *Drug Safety*, *48*(3), 287-303. https://doi.org/10.1007/s40264-024-01499-1

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.-T., Chen, J., Liu, Y., Tang, J., Li, J. y Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, *5*(3), 220-235. https://doi.org/10.1038/s42256-023-00626-4

Dorémus, O., Russon, D., Contrand, B., Guerra-Adames, A., Avalos-Fernandez, M., Gil-Jardiné, C. y Lagarde, E. (2025). Harnessing Moderate-Sized Language Models for Reliable Patient Data Deidentification in Emergency Department Records: Algorithm Development, Validation, and Implementation Study. *JMIR AI*, *4*(1). https://doi.org/10.2196/57828

Elbakary, A., Ben Issaid, C., ElBatt, T. A., Seddik, K. G. y Bennis, M. (2025). MIRA: A Method of Federated Multi-Task Learning for Large Language Models. *IEEE Networking Letters*, *7*(3), 171-175. https://doi.org/10.1109/LNET.2025.3539810

Elmitwalli, S. E. S., Mehegan, J., Gallagher, A. W. A. y Alebshehy, R. (2024). Enhancing sentiment and intent analysis in public health via fine-tuned Large Language Models on tobacco and e-cigarette-related tweets. *Frontiers in Big Data*, *7*. https://doi.org/10.3389/fdata.2024.1501154

Gogineni, K., Suvizi, A. y Venkataramani, G. P. (2025). LLMs on a Budget: System-Level Approaches to Power-Efficient and Scalable Fine-Tuning. *IEEE Open Journal of the Computer Society*, *6*, 987-1000. https://doi.org/10.1109/OJCS.2025.3580498

Gorbatovski, A. V., Razin, A. D., Aliev, A. A. y Kovalchuk, S. V. (2024). Improving question answering in programming domain with pretrained language model finetuning using structured diverse online forum data. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, *24*(6), 1024-1034. https://doi.org/10.17586/2226-1494-2024-24-6-1024-1034

Han, S., Shi, L. y Tsui, F. R. (2025). Enhancing semantical text understanding with fine-tuned large language models: A case study on Quora Question Pair duplicate identification. *PLoS ONE*, *20*(1). https://doi.org/10.1371/journal.pone.0317042

Han, Y., Liu, J., Luo, A., Wang, Y. y Bao, S. (2025). Fine-Tuning LLM-Assisted Chinese Disaster Geospatial Intelligence Extraction and Case Studies. *ISPRS International Journal of Geo-Information*, *14*(2). https://doi.org/10.3390/ijgi14020079

Harris, S., Liu, J., Hadi, H. J., Ahmad, N. y Alshara, M. A. (2025). Benchmarking Hook and Bait Urdu news dataset for domain-agnostic and multilingual fake news detection using large language models. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-98271-x

Hou, Y., Bert, C., Gomaa, A., Lahmer, G., Höfler, D., Weissmann, T., Voigt, R., Schubert, P., Schmitter, C., Depardon, A., Semrau, S., Maier, A. K., Fietkau, R., Huang, Y. y Putz, F. (2024). Fine-tuning a local LLaMA-3 large language model for automated privacy-preserving physician letter generation in radiation oncology. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1493716

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M. y Gelly, S. (2019). Parameter-Efficient Transfer Learning for NLP. https://arxiv.org/abs/1902.00751

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

Huang, W., Zheng, X., Ma, X., Qin, H., Lv, C., Chen, H., Luo, J., Qi, X., Liu, X. y Magno, M. (2024). An empirical study of LLaMA3 quantization: from LLMs to MLLMs. *Visual Intelligence*, *2*(1). https://doi.org/10.1007/s44267-024-00070-x

Iftikhar, S., Alsamhi, S. H. y Davy, S. (2025). Enhancing Sustainability in LLM Training: Leveraging Federated Learning and Parameter-Efficient Fine-Tuning. *IEEE Transactions on Sustainable Computing*, *10*(6), 1158-1172. https://doi.org/10.1109/TSUSC.2025.3592043

Jia, N., Liu, T., Chen, J., Zhang, Y. y Han, S. (2025). Task-Agnostic Adaptive Activation Scaling Network for LLMs. *IEEE Access*, *13*, 31774-31784. https://doi.org/10.1109/ACCESS.2025.3542415

Jung, Y., Hwang, H. y Lee, C. (2025). Prefix Tuning Using Residual Reparameterization. *IEEE Access*, *13*, 54866-54872. https://doi.org/10.1109/ACCESS.2025.3554513

Kadyrbek, N., Tuimebayev, Z. K., Mansurova, M. E. y Viegas, V. M. R. (2025). The Development of Small-Scale Language Models for Low-Resource Languages, with a Focus on Kazakh

and Direct Preference Optimization. *Big Data and Cognitive Computing*, *9*(5). https://doi.org/10.3390/bdcc9050137

Khader, K. A., Hussein, M. S. y Abu-Issa, A. S. (2025). Adapting Large Language Models for Arabic: Comparative Evaluation, Fine-Tuning, and Ethical Deployment. *IEEE Access*, *13*, 182621-182632. https://doi.org/10.1109/ACCESS.2025.3623796

Khoboko, P. W., Marivate, V. N. y Sefara, J. (2025). Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models. *Machine Learning with Applications*, *20*. https://doi.org/10.1016/j.mlwa.2025.100649

Kiziltepe, R. S., Ezin, E., Yentür, Ö., Basbrain, A. M. y Karakus, M. (2025). Advancing Sentiment Analysis for Low-Resource Languages Using Fine-Tuned LLMs: A Case Study of Customer Reviews in Turkish Language. *IEEE Access*, *13*, 77382-77394. https://doi.org/10.1109/ACCESS.2025.3566000

Kuo, S., Tai, S., Lin, H. y Chen, R. (2025). Automated Clinical Trial Data Analysis and Report Generation by Integrating Retrieval-Augmented Generation (RAG) and Large Language Model (LLM) Technologies. *AI (Switzerland)*, *6*(8). https://doi.org/10.3390/ai6080188

Lester, B., Al-Rfou, R. y Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045-3059. https://doi.org/10.18653/v1/2021.emnlp-main.243

Li, H., Zhang, J., Shen, H., Cheng, K. y Huang, X. (2025). KEFT: Knowledge-Enhanced Fine-Tuning for Large Language Models in Domain-Specific Question Answering. *Transactions of the Association for Computational Linguistics*, *13*, 1056-1067. https://doi.org/10.1162/TACL.a.31

Li, L., He, Y., Xu, R., Chen, B., Han, B., Zhao, Y. y Li, J. (2025). Synchronizing LLM-based semantic knowledge bases via secure federated fine-tuning in semantic communication. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1690950

Li, X. L. y Liang, P. (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. En C. Zong, F. Xia, W. Li y R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 4582-4597). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.353

Li, X., He, H., Lu, G., Yue, P., Chen, J., Yang, Z. y Hon, C. (2025). TCM-DS: a large language model for intelligent traditional Chinese medicine edible herbal formulas recommenda-

tions. *Chinese Medicine (United Kingdom)*, *20*(1). https://doi.org/10.1186/s13020-025-01249-0

Lialin, V., Deshpande, V., Yao, X. y Rumshisky, A. (2024). Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning. https://arxiv.org/abs/2303.15647

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 74-81. https://aclanthology.org/W04-1013/

Lin, D., Wen, Y., Wang, W. y Su, Y. (2024). Enhanced Sentiment Intensity Regression Through LoRA Fine-Tuning on Llama 3. *IEEE Access*, *12*, 108072-108087. https://doi.org/10.1109/ACCESS.2024.3438353

Liu, C., Sun, K., Zhou, Q., Duan, Y., Shu, J., Kan, H., Gu, Z. y Hu, J. (2024). CPMI-ChatGLM: parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-56874-w

Liu, J., Song, Y., Jiang, R., Feng, Y., Tao, M. y Li, Y. (2025). Tasks-Embedded Reparameterization: A Novel Framework for Task-Specific Transfer Enhancement With Multitask Prompt Learning. *International Journal of Intelligent Systems*, *2025*(1). https://doi.org/10.1155/int/1688391

Liu, N. y Liu, D. (2025). MPCTF: A Multi-Party Collaborative Training Framework for Large Language Models. *Electronics (Switzerland)*, *14*(16). https://doi.org/10.3390/electronics14163253

Liu, X., Zhu, R., Zha, D., Gao, J., Zhong, S., White, M. y Qiu, M. (2025). Differentially Private Low-Rank Adaptation of Large Language Model Using Federated Learning. *ACM Transactions on Management Information Systems*, *16*(2). https://doi.org/10.1145/3682068

Meng, D., Wang, Z., Phuntsho, T. y Gonsalves, T. (2025). TriPlaNet: Enhancing machine-paraphrasing plagiarism detection through triplet network and contrastive learning. *Egyptian Informatics Journal*, *31*. https://doi.org/10.1016/j.eij.2025.100752

Nowak, S., Wulff, B., Layer, Y. C., Theis, M., Isaak, A., Salam, B., Block, W., Kuetting, D. L. R., Pieper, C. C., Luetkens, J. A., Attenberger, U. I. y Sprinkart, A. M. (2025). Privacy-ensuring Open-weights Large Language Models Are Competitive with Closed-weights GPT-4o in Extracting Chest Radiography Findings from Free-Text Reports. *Radiology*, *314*(1). https://doi.org/10.1148/radiol.240895

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. y Lowe, R. (2022). Training language models to follow instructions with human feedback. https://arxiv.org/abs/2203.02155

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... y Moher, D. (2021). Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas. *Revista Española de Cardiología*, *74*(9), 790-799. https://doi.org/10.1016/j.recesp.2021.06.016

Paiola, P. H., García, G. L., Correia, J. V. M., Manesco, J. R. R., Garcia, A. L. A. y Papa, J. P. (2025). The Bode Family of Large Language Models: Investigating the Frontiers of LLMs in Brazilian Portuguese. *Journal of the Brazilian Computer Society*, *31*(1), 918-939. https://doi.org/10.5753/jbcs.2025.5812

Papineni, K., Roukos, S., Ward, T. y Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. En P. Isabelle, E. Charniak y D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Peng, Y. T. y Lei, C. L. (2025). A Keyword-Guided Training Approach to Large Language Models for Judicial Document Generation. *CMES - Computer Modeling in Engineering and Sciences*, *145*(3), 3969-3992. https://doi.org/10.32604/cmes.2025.073258

Prottasha, N. J., Mahmud, A., Sobuj, M. S. I., Bhat, P., Kowsher, M., Yousefi, N. y Garibay, O. O. (2024). Parameter-efficient fine-tuning of large language models using semantic knowledge tuning. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-75599-4

Qi, H., Chung-Ho, S., Chan, R. Y. Y. y Man Victor Wong, C. (2025). Transforming Applied Behavior Analysis Therapy: An Internet of Things-Guided, Retrieval-Augmented Large Language Model Framework. *IEEE Access*, *13*, 149679-149694. https://doi.org/10.1109/ACCESS.2025.3600316

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S. y Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Thirty-*

seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=HPuSIXJaa9

Rathore, B. S. y Chaurasia, S. (2025). Fine tuning large language models for hate speech detection in Hinglish and code mixed custom dataset through a socially responsible approach for safer digital platforms. *Discover Sustainability*, *6*(1). https://doi.org/10.1007/s43621-025-02190-w

Ren, R., Ma, J. y Luo, J. (2025). Large language model for patent concept generation. *Advanced Engineering Informatics*, *65*. https://doi.org/10.1016/j.aei.2025.103301

Richter-Pechanski, P., Seiferling, M., Kiriakou, C., Schwab, D. M., Geis, N. A., Dieterich, C. y Frank, A. (2025). Medication information extraction using local large language models. *Journal of Biomedical Informatics*, *169*. https://doi.org/10.1016/j.jbi.2025.104898

Rullo, R., Maatouk, A., Huang, T., Chen, J., Qiu, W., O'Connor, G., Womack, J. A., Sadak, T. I., Rodriguez, C., Carneiro, P., de Jesús-Espinosa, T., Marshall, A., Ying, R. y Ramos, S. R. (2025). Interdisciplinary Development and Fine-Tuning of CARDIO, a Large Language Model for Cardiovascular Health Education in HIV Care: Tutorial. *Journal of Medical Internet Research*, *27*(1). https://doi.org/10.2196/77053

Savage, T. R., Ma, S. P., Boukil, A., Rangan, E. S., Patel, V., López, I. y Chen, J. H. (2025). Fine-Tuning Methods for Large Language Models in Clinical Medicine by Supervised Fine-Tuning and Direct Preference Optimization: Comparative Evaluation. *Journal of Medical Internet Research*, *27*. https://doi.org/10.2196/76048

Seo, J.-w. y Jung, H.-y. (2025). Eliminating Packing-Aware Masking via LoRA-Based Supervised Fine-Tuning of Large Language Models. *Mathematics*, *13*(20). https://doi.org/10.3390/math13203344

Shen, A., Lai, Z., Li, D. y Hu, X. (2025). Optimizing Fine-Tuning in Quantized Language Models: An In-Depth Analysis of Key Variables. *Computers, Materials and Continua*, *82*(1), 307-325. https://doi.org/10.32604/cmc.2024.057491

Shen, J., Su, N. y Lin, Y. (2025). Effective Multi-Class Sentiment Analysis Using Fine-Tuned Large Language Model with KNIME Analytics Platform. *Systems*, *13*(7). https://doi.org/10.3390/systems13070523

Sindhujan, A., Kanojia, D. y Orăsan, C. (2025). Reference-Less Evaluation of Machine Translation: Navigating Through the Resource-Scarce Scenarios. *Information (Switzerland)*, *16*(10). https://doi.org/10.3390/info16100916

Song, Y., Lv, C., Zhu, K. y Qiu, X. (2025). LoRA fine-tuning of Llama3 large model for intelligent fishery field. *Discover Computing*, *28*(1). https://doi.org/10.1007/s10791-025-09663-6

Sousa, S., Jantscher, M., Kröll, M. y Kern, R. (2025). Large Language Models for Electronic Health Record De-Identification in English and German. *Information (Switzerland)*, *16*(2). https://doi.org/10.3390/info16020112

Sun, Y., Zhang, C., Wang, C. y Han, Y. (2024). MIRA-ChatGLM: A Fine-Tuned Large Language Model for Intelligent Risk Assessment in Coal Mining. *Applied Sciences (Switzerland)*, *14*(24). https://doi.org/10.3390/app142412072

Sun, Y. y Zhou, T. (2025). DialogueMLLM: Transforming Multimodal Emotion Recognition in Conversation Through Instruction-Tuned MLLM. *IEEE Access*, *13*, 121048-121060. https://doi.org/10.1109/ACCESS.2025.3591447

Tong, X., Ding, X., Jia, H., Yuan, Y., Liu, L., Wang, Y., Xiong, Z., Yang, X., Im, S. K. y Wang, M. H. (2025). XuanHuGPT: parameter-efficient fine-tuning of large language model in the field of traditional Chinese medicine. *Chinese Medicine (United Kingdom)*, *20*(1). https://doi.org/10.1186/s13020-025-01200-3

Tran, H., Yang, Z., Yao, Z. y Yu, H. (2024). BioInstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, *31*(9), 1821-1832. https://doi.org/10.1093/jamia/ocae122

Uluırmak, B. A. y Kurban, R. (2025). Fine-tuning Large Language Models for Turkish Flutter Code Generation. *Sakarya University Journal of Computer and Information Sciences*, *8*(4), 637-650. https://doi.org/10.35377/saucis...1722643

Vithanage, D. L., Deng, C., Wang, L., Yin, M., Alkhalaf, M., Zhang, Z., Zhu, Y. y Yu, P. (2025). Adapting Generative Large Language Models for Information Extraction from Unstructured Electronic Health Records in Residential Aged Care: A Comparative Analysis of Training Approaches. *Journal of Healthcare Informatics Research*, *9*(2), 191-219. https://doi.org/10.1007/s41666-025-00190-z

Wang, Q. y Shen, S. (2025). Activation-Guided Low-Rank Parameter Adaptation for Efficient Model Fine-Tuning. *IEEE Access*, *13*, 70909-70918. https://doi.org/10.1109/ACCESS.2025.3533701

Wang, Y., Chen, B. y Sun, J. (2025). Leveraging Asymmetric Adaptation with Dynamic Sparse LoRA for Enhanced Nuance in LLM-Based Offensive Language Detection. *Symmetry*, *17*(7). https://doi.org/10.3390/sym17071076

Wu, J., Chen, L., Fang, S. y Wu, C. (2024). An Application Programming Interface (API) Sensitive Data Identification Method Based on the Federated Large Language Model. *Applied Sciences (Switzerland)*, *14*(22). https://doi.org/10.3390/app142210162

Wu, W., Xu, D., Liu, L., Wang, B., Zhao, Y., Cheng, X. y Li, X. (2025). Domain-Specialized Large Language Model for Corrosion Analysis: Construction and Evaluation of Corr-Lora-RAG. *Applied Sciences (Switzerland)*, *15*(16). https://doi.org/10.3390/app15169226

Xiang, B. y Shao, Y. (2024). SumLLaMA: Efficient Contrastive Representations and Fine-Tuned Adapters for Bug Report Summarization. *IEEE Access*, *12*, 78562-78571. https://doi.org/10.1109/ACCESS.2024.3397326

Xiong, J., Pan, L., Liu, Y., Zhu, L., Zhang, L. y Tan, S. (2025). Enhancing Plant Protection Knowledge with Large Language Models: A Fine-Tuned Question-Answering System Using LoRA. *Applied Sciences (Switzerland)*, *15*(7). https://doi.org/10.3390/app15073850

Xu, J., Gao, J., Liu, X., Yang, G., Liu, J., Long, Y., Huang, Z. y Yang, K. (2025). CPEL: A Causality-Aware, Parameter-Efficient Learning Framework for Adaptation of Large Language Models with Case Studies in Geriatric Care and Beyond. *Mathematics*, *13*(15). https://doi.org/10.3390/math13152460

Xu, X. O., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J. A., Ghassemi, M., Dey, A. K. y Wang, D. (2024). Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(1). https://doi.org/10.1145/3643540

Yang, G., Liu, X., Shi, J., Wang, Z. y Wang, G. (2024). TCM-GPT: Efficient pre-training of large language models for domain adaptation in Traditional Chinese Medicine. *Computer Methods and Programs in Biomedicine Update*, *6*. https://doi.org/10.1016/j.cmpbup.2024.100158

Yang, W., Jin, C., Chen, G. y Jin, H. (2025). CE-Prompt: Enhance prompt expression stability by multiple understanding. *PeerJ Computer Science*, *11*. https://doi.org/10.7717/peerj-cs.3231

Yang, Y., Li, Z., Li, L., Wang, G., Hu, Y., Peng, X. y Li, Y. (2025). A GRPO-Based Approach to Hierarchical Knowledge Distillation in Military Information Extraction. *IEEE Access*, *13*, 187627-187637. https://doi.org/10.1109/ACCESS.2025.3621013

Ye, H., Xie, X., Xie, F., Zuo, J. y Bu, C. (2025). LoRA-Adv: Boosting Text Classification in Large Language Models Through Adversarial Low-Rank Adaptations. *IEEE Access*, *13*, 103234-103244. https://doi.org/10.1109/ACCESS.2025.3579539

Zain, M., Hussain, N., Qasim, A., Mehak, G., Ahmad, F., Sidorov, G. y Gelbukh, A. F. (2025). RU-OLD: A Comprehensive Analysis of Offensive Language Detection in Roman Urdu Using Hybrid Machine Learning, Deep Learning, and Transformer Models. *Algorithms*, *18*(7). https://doi.org/10.3390/a18070396

Zhang, B., Ma, H., Li, D., Ding, J., Wang, J., Xu, B. y Lin, H. (2025). Efficient Tuning of Large Language Models for Knowledge-Grounded Dialogue Generation. *Transactions of the Association for Computational Linguistics*, *13*, 1007-1031. https://doi.org/10.1162/TACL.a.17

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. y Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. https://arxiv.org/abs/1904.09675

Zhang, Y. (2025). Enhancing Contextual Data Analysis through Retrieval-Augmented Fine-Tuning of Large Language Models. *Informatica (Slovenia)*, *49*(29). https://doi.org/10.31449/inf.v49i29.8094

Zhao, Y., Zhu, X., Tang, W., Zhou, L., Feng, L. y Tang, M. (2025). DAFT: Domain-Augmented Fine-Tuning for Large Language Models in Emotion Recognition of Health Misinformation. *Applied Sciences (Switzerland)*, *15*(23). https://doi.org/10.3390/app152312690

Zhou, L., Chen, Y., Li, X., Li, Y., Li, N., Wang, X. y Zhang, R. (2024). A New Adapter Tuning of Large Language Model for Chinese Medical Named Entity Recognition. *Applied Artificial Intelligence*, *38*(1). https://doi.org/10.1080/08839514.2024.2385268

Zhou, W., Huang, M., Liu, S., You, Q. y Meng, F. (2025). Research on the Construction and Application of Earthquake Emergency Information Knowledge Graph Based on Large Language Models. *IEEE Access*, *13*, 127742-127757. https://doi.org/10.1109/ACCESS.2025.3586370

Zhu, X., Gao, Z. y Wang, X. A. (2025). Full-Parameter Fine-Tuning Method of LLMs for Sports Injury Prevention and Treatment. *International Journal of Mobile Computing and Multimedia Communications*, *16*(1). https://doi.org/10.4018/IJMCMC.376486

Zou, W., Shen, Z., Li, Q., Ge, J., Li, C., Chen, X., Shen, X., Huang, L. y Luo, B. (2025). Experimental Evaluation of Parameter-Efficient Fine-Tuning for Software Engineering

Tasks. *ACM Transactions on Software Engineering and Methodology*, *34*(7). https://doi.org/10.1145/3722107

10. APÉNDICES

APÉNDICE A: CORPUS COMPLETO DE ESTUDIOS INCLUIDOS

Las tablas siguientes presentan los 84 estudios incluidos en la revisión sistemática, organizados por dominio de aplicación. Para cada estudio se reporta la técnica principal de *fine-tuning*, la tarea evaluada, la métrica principal de evaluación declarada por los autores, el valor del *baseline* de referencia y el resultado reportado, junto con el cambio porcentual relativo ($\Delta\,\%$) calculado sobre la misma métrica. Los valores negativos de $\Delta\,\%$ indican que la técnica evaluada no superó al *baseline*.

A.1 SALUD Y BIOMEDICINA (N = 30)

Tabla 10.1

Corpus completo – Salud y Biomedicina (n = 30).

Referencia	Técnica	Tarea	Métrica principal	Baseline	Resultado	Δ %
Full Fine-Tuning						
(Sousa et al., 2025)	Full FT	Desidentificación	—	—	—	—
(X. O. Xu et al., 2024)	IT	Clasif. salud mental	—	—	—	—
PEFT						
(Nowak et al., 2025)	*LoRA*	Extracción clínica	—	—	—	—
(B. Zhang et al., 2025)	*Adapter*	Diálogo con conocimiento	—	—	—	—
(L. Zhou et al., 2024)	*Adapter*	NER médico (chino)	—	—	—	—
(Zhao et al., 2025)	*DAFT*	Desinformación en salud	Accuracy	76,95 %	84,77 %	+10,2 %
(Adhikary et al., 2025)	*LoRA*	Q&A salud menstrual	—	—	—	—
(Bayram et al., 2025)	*LoRA*	Q&A médico (turco)	Similitud	0,4627	0,5151	+11,3 %
(Elmitwalli et al., 2024)	*LoRA*	Análisis de sentimiento	Accuracy	0,65	0,94	+44,6 %
(Khader et al., 2025)	*LoRA*	Q&A médico (árabe)	—	—	—	—

Continúa en la página siguiente

Tabla 10.1 – continuación

Referencia	Técnica	Tarea	Métrica principal	Baseline	Resultado	Δ %
(X. Li et al., 2025)	*LoRA*	Recom. fórmulas TCM	Precision	0,3848	0,9924	+157,9 %
(Qi et al., 2025)	*LoRA*	Diálogo terapia ABA	—	—	—	—
(Richter-Pechanski et al., 2025)	*LoRA*	Extracción medicamentos	—	—	—	—
(Tong et al., 2025)	*LoRA*	Q&A medicina TCM	—	—	—	—
(Tran et al., 2024)	*LoRA*	Q&A biomédico	—	—	—	—
(Vithanage et al., 2025)	*LoRA*	NER clínico	—	—	—	—
(J. Xu et al., 2025)	*LoRA*	Diálogo geriátrico	ROUGE	8,60 %	34,96 %	+306,5 %
(G. Yang et al., 2024)	*LoRA*	Medicina tradicional	Accuracy	0,247	0,290	+17,4 %
(Bao et al., 2025)	*LoRA*	Diálogo terapéutico	—	—	—	—
(C. Liu et al., 2024)	P-Tuning v2	Prescripción TCM	—	—	—	—
(W. Yang et al., 2025)	*Prompt Tuning*	Q&A médico (KG)	—	—	—	—
(Bui et al., 2025)	*LoRA*	Q&A médico (vietnamita)	—	—	—	—
(Chao et al., 2025)	*QLoRA*	Resumen clínico	—	—	—	—
(Dietrich y Hollstein, 2025)	*QLoRA*	NER farmacovigilancia	—	—	—	—
(Dorémus et al., 2025)	*QLoRA*	Desidentificación	—	—	—	—
(Hou et al., 2024)	*QLoRA*	Gen. cartas médicas	ROUGE-1	0,161	0,352	+118,6 %
Múltiples						
(Zhu et al., 2025)	Full FT (GaLore)	Prev. lesiones deportivas	BLEU-4	29,415	29,589	+0,6 %
(Rullo et al., 2025)	*LoRA*+GRPO	Educación cardiovascular (HIV)	Human Eval. Accuracy	4,16	5,0	+20,2 %
(Kuo et al., 2025)	*LoRA*+GRPO	Reportes clínicos	—	—	—	—
(Savage et al., 2025)	SFT+DPO	Razonamiento clínico	—	—	—	—

Nota. IT = *Instruction Tuning*; Full FT = *Full Fine-Tuning*; SFT = *Supervised Fine-Tuning*; DPO = *Direct Preference Optimization*; GRPO = *Group Relative Policy Optimization*; DAFT = *Domain-Augmented Fine-Tuning*; GaLore = *Gradient Low-Rank Projection*; TCM = *Traditional Chinese Medicine* (medicina tradicional china); KG = *Knowledge*

Graph (grafo de conocimiento); NER = *Named Entity Recognition* (reconocimiento de entidades nombradas); ABA = *Applied Behavior Analysis* (análisis conductual aplicado); Human Eval. Accuracy = exactitud evaluada por jueces humanos en escala numérica definida por cada estudio (escala 1-5 en (Rullo et al., 2025)). Los guiones (—) en las columnas Baseline, Resultado y Δ% indican que el estudio no reporta un par de comparación homogéneo que permita calcular una mejora relativa, ya sea porque evalúa múltiples técnicas o modelos sin un *baseline* único de referencia, o porque los resultados no son directamente comparables entre sí.

A.2 LENGUAJE Y HUMANIDADES (N = 21)

Tabla 10.2

Corpus completo – Lenguaje y Humanidades (n = 21).

Referencia	Técnica	Tarea	Métrica principal	Baseline	Resultado	Δ%
PEFT						
(K. Chen et al., 2025)	*Adapter*	Clasificación GLUE	Promedio	86,4	89,0	+3,0 %
(Anghel et al., 2025)	*LoRA*	Evaluación educativa	Human Eval. Accuracy	7,94	8,21	+3,4 %
(Z. Chen y Liu, 2025)	*LoRA*	Q&A patrimonio cultural	ROUGE-L	22,39 %	47,85 %	+113,7 %
(S. Han et al., 2025)	*LoRA*	Detección de duplicados	F1	82,02 %	84,90 %	+3,5 %
(Harris et al., 2025)	*LoRA*	Fake news (urdu)	—	—	—	—
(Kiziltepe et al., 2025)	*LoRA*	Sentimiento (turco)	—	—	—	—
(H. Li et al., 2025)	*LoRA*	Q&A general	—	—	—	—
(Meng et al., 2025)	*LoRA*	Detección de plagio	—	—	—	—
(Rathore y Chaurasia, 2025)	*LoRA*	Discurso de odio	—	—	—	—
(Y. Wang et al., 2025)	DS-LoRA	Lenguaje ofensivo	Macro-F1	80,1	81,3	+1,5 %
(Ye et al., 2025)	*LoRA*	Texto ofensivo	—	—	—	—
(Zain et al., 2025)	*LoRA*	Lenguaje ofensivo (urdu)	—	—	—	—
(Jia et al., 2025)	Multi-PEFT	Clasif. multi-tarea	Macro-F1	0,8712	0,8736	+0,3 %
(Cheng y Cao, 2025)	*Prompt Tuning*	KGQA multidominio	—	—	—	—

Continúa en la página siguiente

Tabla 10.2 – continuación

Referencia	Técnica	Tarea	Métrica principal	Baseline	Resultado	Δ %
(Huang et al., 2024)	*QLoRA*	Cuantización LLMs	—	—	—	—
(Khoboko et al., 2025)	*QLoRA*	Traducción inglés-zulú	F1	10,52	42	+299,2 %
(Paiola et al., 2025)	*QLoRA*	Benchmarks port. bras.	—	—	—	—
(Sindhujan et al., 2025)	*LoRA*	Quality Estimation (QE) sin referencia	—	—	—	—
Alineación						
(Kadyrbek et al., 2025)	DPO	Q&A en kazajo	Accuracy	56,5 %	62,7 %	+11,0 %
(J. Shen et al., 2025)	ORPO	Análisis de sentimiento	Accuracy	75,33 %	86,67 %	+15,1 %
Múltiples						
(Sun y Zhou, 2025)	*LoRA*+IT	Reconoc. de emociones	—	—	—	—

Nota. IT = *Instruction Tuning*; DPO = *Direct Preference Optimization*; ORPO = *Odds Ratio Preference Optimization*; DS-LoRA = variante de *LoRA* propuesta por los autores; Multi-PEFT = combinación de múltiples métodos *PEFT*; GLUE = *General Language Understanding Evaluation*; KGQA = *Knowledge Graph Question Answering*; QE = *Quality Estimation* (estimación de calidad de traducción sin referencia); Human Eval. Accuracy = exactitud evaluada por jueces humanos en escala numérica definida por cada estudio (escala 0-10 en (Anghel et al., 2025)). Los guiones (—) en las columnas Baseline, Resultado y Δ % indican que el estudio no reporta un par de comparación homogéneo que permita calcular una mejora relativa, ya sea porque evalúa múltiples técnicas o modelos sin un *baseline* único de referencia, o porque los resultados no son directamente comparables entre sí.

A.3 CIENCIAS EXACTAS Y COMPUTACIÓN (N = 20)

Tabla 10.3

Corpus completo – Ciencias Exactas y Computación (n = 20).

Referencia	Técnica	Tarea	Métrica principal	Baseline	Resultado	Δ %
PEFT						
(X. Liu et al., 2025)	DP-LoRA	PEFT federado privado	—	—	—	—

Continúa en la página siguiente

Tabla 10.3 – continuación

Referencia	**Técnica**	**Tarea**	**Métrica principal**	**Baseline**	**Resultado**	**Δ %**
(Ao et al., 2025)	SPLoRA	Seguridad en *fine-tuning*	—	—	—	—
(Y. Chen et al., 2025)	*LoRA*	Razon. sobre grafos	—	—	—	—
(Elbakary et al., 2025)	*LoRA*	Aprendizaje federado	—	—	—	—
(Gogineni et al., 2025)	*LoRA*	Optimización de eficiencia en *fine-tuning* (GPU)	—	—	—	—
(Iftikhar et al., 2025)	*LoRA*	Sentimiento (FL)	Accuracy	71,2 %	73,8 %	+3,7 %
(L. Li et al., 2025)	*LoRA*	PEFT federado	Avg. Acc.	76,6 %	75,4 %	-1,6 %
(N. Liu y Liu, 2025)	*LoRA*	Entren. multi-parte	Accuracy	14,25	65,43	+359,2 %
(Seo y Jung, 2025)	*LoRA*	Reducción de overhead en SFT con packing	—	—	—	—
(Q. Wang y Shen, 2025)	*LoRA*	Razon. sentido común	—	—	—	—
(J. Wu et al., 2024)	*LoRA*	Clasif. datos sensibles	—	—	—	—
(Xiang y Shao, 2024)	*LoRA*	Resumen bug reports	ROUGE-L	39,69	41,01	+3,3 %
(Prottasha et al., 2024)	Múltiples	Clasificación de texto y NLU multi-tarea	—	—	—	—
(Zou et al., 2025)	Múltiples	Ing. de software	—	—	—	—
(Jung et al., 2025)	*Prefix Tuning*	Clasificación GLUE	—	—	—	—
(J. Liu et al., 2025)	*Prompt Tuning*+*LoRA*	Razon. SuperGLUE	—	—	—	—
(Choi y Yang, 2025)	*QLoRA*	Resumen bug reports	ROUGE-1 Recall	61,0 %	84,6 %	+38,7 %
(A. Shen et al., 2025)	*QLoRA*	Análisis de impacto de hiperparámetros en QLoRA	—	—	—	—
(Uluırmak y Kurban, 2025)	*QLoRA*	Gen. código (Flutter)	—	—	—	—
Múltiples						
(Gorbatovski et al., 2024)	*LoRA*+RLHF	Q&A Stack Overflow	ROUGE-1	0,1889	0,1978	+4,7 %

Nota. SFT = *Supervised Fine-Tuning*; RLHF = *Reinforcement Learning from Human Feedback*; DP-LoRA = *Differentially Private LoRA*; SPLoRA = variante de *LoRA* aplicada a la seguridad del ajuste fino; FL = *Federated Learning* (aprendizaje federado); GLUE/SuperGLUE = *benchmarks* de comprensión del lenguaje natural; Avg. Acc. = *Average Accuracy* (exactitud promedio); Gen. = generación; Razon. = razonamiento; Entren. = entrenamiento; Ing. = ingeniería. Los guiones (—) en las columnas Baseline, Resultado y Δ % indican que el estudio no reporta un par de comparación homogéneo que permita calcular una mejora relativa, ya sea porque evalúa múltiples técnicas o modelos sin un *baseline* único de referencia, o porque los resultados no son directamente comparables entre sí.

A.4 INDUSTRIA Y NEGOCIOS (N = 13)

Tabla 10.4

Corpus completo – Industria y Negocios (n = 13).

Referencia	Técnica	Tarea	Métrica principal	Baseline	Resultado	Δ %
PEFT						
(Chitrakala y Pavithra, 2025)	*LoRA*	Gen. redes sociales	—	—	—	—
(Y. Han et al., 2025)	*LoRA*	Extracción geoespacial	—	—	—	—
(D. Lin et al., 2024)	*LoRA*	Análisis de sentimiento	—	—	—	—
(Song et al., 2025)	*LoRA*	Q&A acuicultura	ROUGE-L	0,26	0,35	+34,6 %
(Sun et al., 2024)	*LoRA*	Riesgo minero	—	—	—	—
(W. Wu et al., 2025)	*LoRA*	Riesgo de corrosión	—	—	—	—
(Y. Yang et al., 2025)	*LoRA*	IE textos militares	Macro-F1	0,2140	0,7020	+228,0 %
(Y. Zhang, 2025)	*LoRA*	Q&A turístico	—	—	—	—
(W. Zhou et al., 2025)	*LoRA*	NER emergencias	F1	0,7874	0,8920	+13,3 %
(Cao et al., 2025)	MoELoRA	Mantenimiento minero	ROUGE-L	0,285	0,416	+46,0 %
Múltiples						
(Peng y Lei, 2025)	*LoRA*+RLHF-PPO	Gen. texto legal	—	—	—	—
(Xiong et al., 2025)	*LoRA*+RAG	Q&A protección plantas	ROUGE-2	0,121	0,321	+165,3 %
(Ren et al., 2025)	*LoRA*+RLHF	Gen. de patentes	BLEU-4	11,663	45,360	+288,9 %

Nota. MoELoRA = *Mixture of Experts LoRA*; RLHF = *Reinforcement Learning from Human Feedback*; RLHF-PPO =

RLHF con *Proximal Policy Optimization*; RAG = *Retrieval-Augmented Generation*; IE = extracción de información; Gen. = generación. Los guiones (—) en las columnas Baseline, Resultado y Δ % indican que el estudio no reporta un par de comparación homogéneo que permita calcular una mejora relativa, ya sea porque evalúa múltiples técnicas o modelos sin un *baseline* único de referencia, o porque los resultados no son directamente comparables entre sí.