REVISTA CUBANA
DE TRANSFORMACIÓN DIGITAL

RNPS 2487 • ISSN 2708-3411
Vol. 7 • 2026 • e303

ARTÍCULO ORIGINAL

Estrategias de ingeniería para el ajuste fino supervisado y eficiente de los LLM específicos de dominio: un estudio de caso sobre patrimonio cultural

Engineering strategies for efficient supervised fine-tuning of domain-specific LLMS: a cultural heritage case study

Reynier Hernández Palacios
reynier.hernandez@uic.cu • *https://orcid.org/0009-0004-7746-5495*

UNIÓN DE INFORMÁTICOS DE CUBA

Reynaldo Alonso Reyes
reynaldo.alonso@uic.cu • *https://orcid.org/0000-0002-8568-2041*

UNIVERSIDAD DE CAMAGÜEY IGNACIO AGRAMONTE LOYNAZ

Recibido: 2026-02-25 • Aceptado: 2026-04-16

RESUMEN

Los modelos de lenguaje de gran tamaño (LLMs) muestran un alto rendimiento en tareas generales, pero su adaptación a dominios altamente especializados sigue siendo un desafío bajo restricciones computacionales. Este trabajo presenta un estudio de caso metodológico de supervised fine-tuning (SFT) del modelo Qwen 1.5 de 7B parámetros para tareas de pregunta–respuesta en un dominio especializado, utilizando el patrimonio cultural como caso de validación. Se propone un pipeline eficiente en recursos que combina técnicas de adaptación eficiente de parámetros y optimización computacional, incluyendo DoRA, QLoRA con cuantización de baja precisión, Flash Attention y gradient checkpointing, permitiendo el entrenamiento en hardware de consumo (NVIDIA RTX 3090, 24 GB de VRAM). El modelo fue ajustado sobre un dataset de 1,000 a 4,000 pares de preguntas y respuestas, generado de forma asistida por LLMs y validado mediante revisión experta. La evaluación cualitativa sistemática mostró mejoras claras en contextualización específica del dominio y reducción de errores respecto al modelo base. Los resultados demuestran la viabilidad de especializar LLMs en dominios altamente contextuales bajo entornos de recursos limitados y aportan un marco metodológico replicable para proyectos de adaptación de dominio.

Palabras clave: Large Language Models (LLMs), Supervised Fine-Tuning, Domain Adaptation, Parameter-Efficient Fine-Tuning, Question Answering Systems.

ABSTRACT

Large Language Models (LLMs) achieve strong performance on general-purpose tasks, yet their adaptation to highly specialized domains remains challenging under computational constraints. This work presents a methodological case study on supervised fine-tuning (SFT) of the Qwen 1.5 7B model for question–answering tasks in a specialized domain, using cultural heritage as a validation case. We propose a resource-efficient fine-tuning pipeline that combines parameter-efficient adaptation techniques and computational optimizations, including DoRA, QLoRA with low-bit quantization, Flash Attention, and gradient checkpointing, enabling training on consumer-grade hardware (NVIDIA RTX 3090, 24 GB VRAM). The model was fine-tuned on a dataset of 1,000 to 4,000 question–answer pairs, generated through LLM-assisted workflows and validated via expert review. Systematic qualitative evaluation demonstrated clear improvements in domain-specific contextualization and error reduction compared to the base model. These results demonstrate the feasibility of specializing LLMs for highly contextual domains under resource-constrained environments and provide a replicable methodological framework for domain adaptation projects.

Keywords: Large Language Models (LLMs), Supervised Fine-Tuning, Domain Adaptation, Parameter-Efficient Fine-Tuning, Question Answering Systems.

INTRODUCCIÓN

Los modelos de lenguaje de gran tamaño han transformado el panorama del procesamiento del lenguaje natural al demostrar capacidades emergentes en comprensión semántica, generación de texto y razonamiento contextual (Brown et al., 2020; Zhao et al., 2023). Estos modelos, entrenados sobre corpus masivos y heterogéneos, exhiben un alto grado de generalización, lo que los hace particularmente atractivos para aplicaciones de propósito general. Sin embargo, esta misma generalidad limita su desempeño en dominios que requieren un conocimiento contextual profundo y especializado, tales como el ámbito legal, médico, técnico o patrimonial. Los dominios altamente especializados se caracterizan por la presencia de marcos conceptuales particulares, terminología técnica, contextos implícitos y conocimiento experto que rara vez se encuentra representado de manera suficiente en los datos de pre-entrenamiento de los LLMs. En consecuencia, los modelos generalistas tienden a producir respuestas genéricas o conceptualmente imprecisas, lo que resulta problemático en aplicaciones especializadas donde la fidelidad contextual y la precisión técnica son fundamentales (Gururangan et al., 2020).

El supervised fine-tuning se ha consolidado como un mecanismo eficaz para adaptar LLMs a dominios específicos mediante el uso de datasets etiquetados de alta calidad (Raffel et al., 2020). No obstante, el ajuste supervisado de modelos con miles de millones de parámetros plantea desafíos computacionales considerables, limitando su adopción en entornos académicos o institucionales con infraestructuras modestas. En respuesta a estas limitaciones, han emergido técnicas de adaptación eficiente de parámetros (Parameter-Efficient Fine-Tuning, PEFT), como LoRA, QLoRA y DoRA, que permiten ajustar modelos de gran escala actualizando únicamente una fracción de

sus parámetros entrenables (Hu et al., 2022; Dettmers et al., 2023; Liu et al., 2024). Estas técnicas, combinadas con optimizaciones modernas de memoria y cómputo, hacen viable el fine-tuning de LLMs en hardware de consumo.

Este trabajo presenta un estudio de caso metodológico que aborda el ajuste supervisado eficiente del modelo Qwen 1.5 7B para tareas de QA en dominios altamente especializados, bajo restricciones computacionales realistas. Aunque el enfoque propuesto es generalizable a diversos dominios especializados, este estudio se centra en el patrimonio cultural como caso de validación representativo, dada su naturaleza contextualmente exigente y la necesidad de fidelidad histórica y semántica. A diferencia de enfoques centrados exclusivamente en métricas cuantitativas, este estudio prioriza una validación funcional y cualitativa sistemática, particularmente apropiada para dominios donde las métricas automáticas resultan insuficientes para capturar la calidad contextual del contenido generado. Aunque este estudio se localiza en el patrimonio cultural, el enfoque propuesto no es específico de este dominio y puede transferirse a otros escenarios de adaptación especializada que compartan características similares de alta contextualidad y sensibilidad a errores de precisión.

A partir de las consideraciones anteriores, este trabajo formula la siguiente hipótesis de investigación: la combinación de técnicas de ajuste fino eficiente en parámetros (en particular DoRA y QLoRA) junto con optimizaciones computacionales modernas (incluyendo Flash Attention y gradient checkpointing) permite la especialización efectiva de un modelo de lenguaje de 7 mil millones de parámetros en un dominio altamente especializado utilizando hardware de consumo, produciendo mejoras cualitativas medibles en la precisión contextual y la corrección factual en tareas de pregunta–respuesta específicas del dominio.

Antecedentes

El ajuste fino de modelos de lenguaje ha sido ampliamente estudiado como estrategia para la adaptación de modelos preentrenados a tareas y dominios específicos. Estudios previos han demostrado que el supervised fine-tuning permite refinar el comportamiento de los modelos sin requerir un entrenamiento desde cero, reduciendo significativamente los costos computacionales y de datos (Raffel et al., 2020; Gururangan et al., 2020). En el ámbito de las técnicas PEFT, LoRA introdujo la idea de congelar los pesos originales del modelo y entrenar únicamente matrices de bajo rango insertadas en capas específicas, reduciendo drásticamente el número de parámetros entrenables (Hu et al., 2022). QLoRA extendió este enfoque mediante la cuantización del modelo base a 4 bits, permitiendo el ajuste de modelos de gran tamaño en GPUs con memoria limitada sin comprometer la estabilidad del entrenamiento (Dettmers et al., 2023). Más recientemente, DoRA propuso una descomposición de las actualizaciones de peso en componentes de magnitud y dirección, mejorando la eficiencia de adaptación y la estabilidad numérica (Liu et al., 2024). Paralelamente, se han desarrollado optimizaciones a nivel de atención y ejecución, como Flash Attention, que reducen el consumo de memoria y aceleran el cálculo del mecanismo de atención en arquitecturas Transformer (Dao et al., 2022). Técnicas como el gradient checkpointing y la compilación JIT del grafo computacional también han demostrado ser efectivas para el entrenamiento eficiente de modelos de gran escala (Chen et al., 2016).

En el contexto de dominios especializados, diversos trabajos han explorado el uso de LLMs para tareas específicas en campos como las humanidades digitales, el patrimonio cultural, el derecho, la medicina y la ingeniería técnica (Gururangan et al., 2020; Kestemont et al., 2022). Sin embargo, la mayoría de estos estudios se apoyan en modelos de propósito general o en ajustes limitados, sin abordar de manera explícita los desafíos de eficiencia computacional ni la replicabilidad en entornos con recursos restringidos. Este trabajo se posiciona en la intersección entre la ingeniería eficiente de LLMs y su aplicación a dominios altamente especializados, aportando una perspectiva metodológica orientada a la implementación realista y sostenible, validada mediante un caso de estudio patrimonial exigente.

METODOLOGÍA

Esta sección describe el pipeline completo de supervised fine-tuning empleado para adaptar el modelo Qwen 1.5 7B a un dominio altamente especializado, enfatizando las decisiones de ingeniería orientadas a maximizar la eficiencia computacional sin comprometer la calidad de la adaptación. El enfoque metodológico presentado es generalizable a diversos dominios especializados que requieran alta contextualización y precisión conceptual. La Figura 1 ilustra el pipeline metodológico general seguido en este estudio, desde la construcción del dataset hasta la validación cualitativa del modelo ajustado. Este flujo de trabajo fue diseñado para garantizar transparencia metodológica y replicabilidad en entornos con restricciones computacionales realistas.

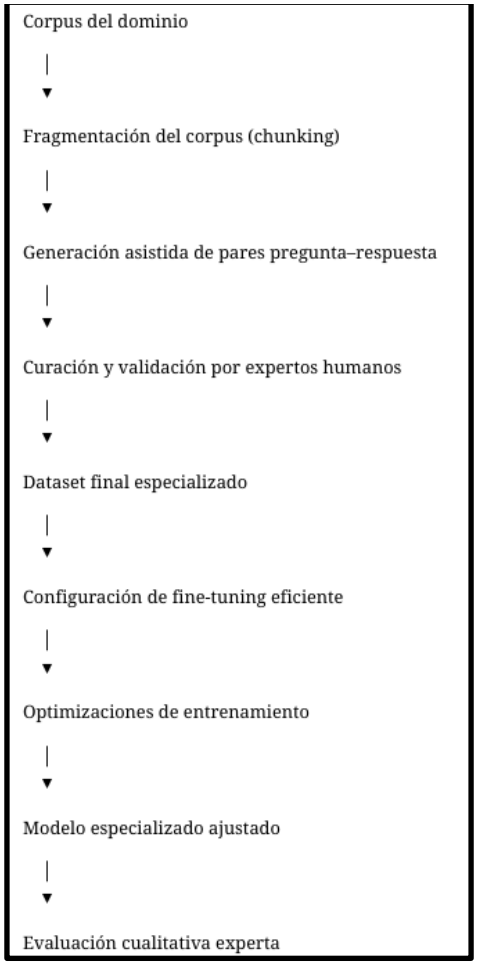

Figura 1. Pipeline para el ajuste fino eficiente en dominios especializados.

Modelo base

El modelo seleccionado como base es Qwen 1.5 7B, un LLM con aproximadamente siete mil millones de parámetros, diseñado para ofrecer un equilibrio entre capacidad de razonamiento y eficiencia computacional. A pesar de su tamaño, este modelo puede considerarse un Small Language Model (SLM) relativo dentro del ecosistema actual de LLMs, lo que lo convierte en un candidato adecuado para escenarios de ajuste fino con hardware de consumo. El entrenamiento se llevó a cabo bajo restricciones computacionales realistas, utilizando una GPU NVIDIA RTX 3090 con 24GB de VRAM, acompañada por un sistema con memoria RAM suficiente para soportar el preprocesamiento del dataset. Estas limitaciones motivaron la adopción de técnicas avanzadas de optimización de memoria y cómputo descritas en las subsecciones siguientes.

Estrategia de Supervised fine-tuning

El ajuste supervisado se formuló como una tarea de aprendizaje sobre pares estructurados de instrucción—respuesta, donde cada ejemplo del dataset guía explícitamente el comportamiento esperado del modelo en el dominio especializado. Este enfoque permite reforzar patrones lingüísticos, semánticos y contextuales específicos, reduciendo ambigüedades presentes en modelos de propósito general. La estrategia es aplicable a cualquier dominio que requiera adaptación contextual profunda. Para habilitar el fine-tuning bajo restricciones de memoria, se integraron múltiples técnicas de Parameter-Efficient Fine-Tuning (PEFT).

DoRA Y QLoRA con cuantización de baja precisión

DoRA extiende el enfoque LoRA tradicional al descomponer las actualizaciones de los pesos en componentes de magnitud y dirección. Esta separación permite capturar de manera más eficiente las transformaciones necesarias para la adaptación al dominio, reduciendo el número de parámetros entrenables sin degradar el rendimiento.

La configuración específica de DoRA empleada fue:

- Rank: 16 para las capas de atención y 8 para las capas feed-forward
- Alpha: 32
- Capas objetivo: Se aplicó DoRA a las proyecciones de query, key, value y output en todas las capas de atención, así como a las proyecciones up y down de los bloques feed-forward.
- Dropout: 0.05.

Tabla 1: Configuración de hiperparámetros para DoRA y QLoRA

Parámetro	Valor	Descripción
Rank (capas atención)	16	Rango de las matrices de bajo rango en capas de atención
Rank (capas feed-forward)	8	Rango de las matrices de bajo rango en capas feed-forward
Alpha	32	Factor de escalado para las adaptaciones
Dropout	0.05	Tasa de dropout aplicada durante el entrenamiento
Esquema de cuantización	NF4 (4 bits)	Esquema de cuantización de baja precisión aplicada al modelo base
Parámetros entrenables	~42 millones	Aproximadamente 0.6% del total de parámetros del modelo

Esta configuración resultó en aproximadamente 42 millones de parámetros entrenables (menos del 0.6% del total del modelo), permitiendo un ajuste eficiente mientras se mantenía capacidad expresiva suficiente para la adaptación al dominio. Su adopción fue motivada por la necesidad de maximizar la eficiencia de VRAM manteniendo estabilidad numérica durante el entrenamiento.

El modelo base fue cuantizado a 4 bits utilizando el esquema de cuantización NF4 (Normal Float 4), optimizado para preservar la distribución de pesos de modelos preentrenados. Esta combinación reduce drásticamente la huella de memoria del modelo base a aproximadamente 3.8GB (en NF4 cuantizado), permitiendo que el ajuste fino se realice en GPUs con 24GB de VRAM sin sacrificar la estabilidad del proceso de entrenamiento. La memoria restante se utiliza para almacenar activaciones, gradientes de los adaptadores y estados del optimizador.

Optimización del mecanismo de atención

Se empleó Flash Attention, una implementación altamente optimizada del mecanismo de atención, diseñada para minimizar accesos a memoria y acelerar el cálculo de atención en arquitecturas NVIDIA modernas. Esta técnica resultó clave para reducir el consumo de VRAM y el tiempo de entrenamiento por época. Flash Attention implementa el cálculo de atención utilizando: Tiling de operaciones a nivel de bloques para maximizar reuso de caché, fusión de kernels para reducir lecturas/escrituras a memoria global y optimizaciones específicas para arquitecturas Ampere y posteriores. En la práctica, Flash Attention redujo el consumo de memoria de las operaciones de atención en aproximadamente 40-50% y aceleró el cálculo de atención en aproximadamente 2-3x comparado con implementaciones estándar, resultando crítico para manejar secuencias largas dentro de las restricciones de VRAM.

Optimización del proceso de entrenamiento

Se aplicaron técnicas adicionales para equilibrar memoria y velocidad:

1. Gradient checkpointing selectivo: Activado en todas las capas del modelo para reducir la memoria requerida para almacenar activaciones intermedias. Aunque esto incrementa el tiempo de entrenamiento en aproximadamente 20%, permitió utilizar batch sizes mayores, resultando en una mejor eficiencia global y convergencia más estable.

2. Compilación del modelo con torch.compile: El modelo fue compilado utilizando PyTorch 2.0+ con torch.compile en modo reduce-overhead, que optimiza el grafo computacional mediante compilación JIT. Esto aceleró las iteraciones de entrenamiento en aproximadamente 15-20% después del costo inicial de compilación.

3. Gradient accumulation: Se empleó acumulación de gradientes con 4 pasos de acumulación, permitiendo simular batch sizes efectivos de mayor tamaño sin exceder las limitaciones de VRAM.

Tabla 2: Configuración de hardware y optimizaciones de entrenamiento

Componente / Técnica	Especificación / Configuración	Impacto
GPU	NVIDIA RTX 3090 (24 GB VRAM)	Hardware de consumo utilizado para el entrenamiento
Batch size efectivo	Simulado mediante gradient accumulation (4 pasos)	Permite mayor tamaño de lote sin exceder memoria

Componente / Técnica	Especificación / Configuración	Impacto
Técnica de atención	Flash Attention	Reducción significativa del consumo de memoria y aceleración del cálculo de atención

Configuración experimental y validación

La evaluación del modelo ajustado se diseñó considerando las limitaciones inherentes al contexto del proyecto, particularmente la ausencia de métricas cuantitativas exhaustivas en dominios donde la evaluación automática resulta insuficiente para capturar la calidad contextual. En consecuencia, se priorizó una validación funcional y cualitativa sistemática, metodología particularmente apropiada para dominios especializados donde la precisión conceptual y la coherencia contextual son más relevantes que métricas superficiales de similitud textual. El proceso de supervised fine-tuning se ejecutó en un entorno de hardware de consumo compuesto por una GPU NVIDIA RTX 3090 con 24GB de VRAM, suficiente para soportar el ajuste del modelo Qwen 1.5 7B mediante técnicas de adaptación eficiente de parámetros. El sistema contó con memoria RAM adecuada para el manejo del dataset y almacenamiento SSD para optimizar el acceso a datos durante el entrenamiento. La configuración de entrenamiento se orientó a maximizar la eficiencia del uso de memoria y cómputo, priorizando la estabilidad del proceso sobre la exploración exhaustiva de hiperparámetros. Dado el carácter metodológico del estudio, el énfasis se colocó en la viabilidad técnica del pipeline propuesto para adaptación de dominio en general.

Estrategia de validación

La validación del modelo ajustado se llevó a cabo mediante dos mecanismos complementarios:

- Demostración funcional: El modelo resultante fue integrado en un entorno de prueba que permitía realizar consultas de pregunta—respuesta relacionadas con el dominio especializado. Este entorno facilitó la comparación directa entre las respuestas generadas por el modelo base y el modelo ajustado, utilizando una interfaz web simple que presentaba ambas respuestas de forma anónima para reducir sesgos de evaluación.

- Evaluación cualitativa experta sistemática: Las respuestas generadas fueron evaluadas de forma cualitativa mediante un protocolo estructurado que garantizó sistematicidad y rigor metodológico. El protocolo de evaluación se diseñó específicamente para capturar dimensiones de calidad relevantes en dominios especializados que las métricas automáticas no pueden medir adecuadamente.

Posteriormente, su comparación con el modelo base se realizó de forma directa, utilizando el modelo base Qwen 1.5 7B sin ajuste como referencia. Las diferencias observadas se centraron principalmente en la calidad contextual, la precisión conceptual y la adecuación al dominio especializado de las respuestas, más que en aspectos puramente lingüísticos, los cuales ya se encontraban bien resueltos en el modelo preentrenado. La selección del modelo base sin ajuste como único baseline se justifica por el objetivo metodológico del estudio: demostrar la viabilidad y efectividad del pipeline de ajuste propuesto. Comparaciones con otros modelos ajustados o con modelos de mayor tamaño quedan fuera del alcance de este estudio de caso, aunque constituyen direcciones valiosas para trabajo futuro.

Protocolo de evaluación cualitativa

Dada la complejidad contextual del dominio objetivo, la evaluación del modelo ajustado se realizó mediante un protocolo cualitativo estructurado diseñado para capturar dimensiones de calidad de respuesta que no pueden ser evaluadas adecuadamente mediante métricas automáticas tradicionales. La evaluación fue realizada por tres expertos independientes con experiencia en estudios de patrimonio cultural y análisis académico del dominio. Para minimizar posibles sesgos en la evaluación, las respuestas generadas por el modelo base y por el modelo ajustado fueron presentadas de forma anónima mediante una interfaz web simple, sin indicar cuál de los sistemas había producido cada respuesta.

Para el proceso de evaluación se seleccionaron aleatoriamente 100 preguntas a partir de un subconjunto de evaluación reservado del dataset. Para cada pregunta, los evaluadores analizaron las respuestas generadas tanto por el modelo base como por el modelo ajustado. Cada respuesta fue evaluada utilizando una escala tipo Likert de 5 puntos en tres dimensiones principales de calidad:

- Contextualización específica del dominio: grado en que la respuesta demuestra comprensión adecuada del contexto cultural e histórico de la pregunta.
- Corrección factual y conceptual: grado en que la información proporcionada es precisa y consistente con el conocimiento experto del dominio.
- Preferencia global de la respuesta: valoración general del evaluador sobre cuál de las respuestas aborda mejor la pregunta planteada.

Las puntuaciones otorgadas por los evaluadores fueron agregadas calculando la media de las evaluaciones para cada dimensión a través de todos los evaluadores y preguntas. Los porcentajes de mejora reportados en la Tabla 2 fueron calculados como el incremento relativo entre las puntuaciones medias del modelo base y del modelo ajustado. Con el fin de evaluar la consistencia de las valoraciones cualitativas, se calculó una medida de acuerdo entre evaluadores utilizando el coeficiente Kappa de Cohen aplicado a las decisiones de preferencia entre modelos. El coeficiente Kappa obtenido fue de 0.71, indicando un nivel de acuerdo sustancial entre evaluadores según la clasificación de Landis y Koch (Liu et al., 2024), lo que respalda la fiabilidad del proceso de evaluación cualitativa. Aunque este tipo de evaluación no proporciona la misma granularidad estadística que los benchmarks automáticos a gran escala, resulta particularmente apropiado para dominios especializados donde la precisión contextual profunda y la coherencia semántica son más relevantes que la mera similitud textual superficial.

RESULTADOS Y DISCUSIÓN

Construcción del dataset

La calidad del dataset es un factor determinante en el éxito del supervised fine-tuning de modelos de lenguaje, particularmente en dominios especializados donde el contexto específico, la terminología técnica y la precisión conceptual desempeñan un rol central. En este estudio, el dataset fue diseñado específicamente para tareas de pregunta—respuesta (QA) orientadas a dominios altamente contextuales, utilizando contenidos del patrimonio cultural como caso representativo de un dominio especializado exigente.

Dominio y alcance del dataset

El dominio objetivo del dataset corresponde a un área especializada que requiere alta contextualización y conocimiento implícito: el patrimonio cultural. Este dominio fue seleccionado como caso de validación por su

naturaleza contextualmente exigente, que lo convierte en un proxy representativo para evaluar la capacidad de adaptación del modelo a escenarios especializados. Los datos incluyen referencias a eventos históricos, contextos socioculturales específicos, expresiones idiomáticas patrimoniales y conocimientos técnicos del dominio, diseñados para evaluar y reforzar la capacidad del modelo de generar respuestas precisas y contextualmente fundamentadas. El tamaño final del dataset se encuentra en un rango aproximado de 1,000 a 4,000 pares Q&A, una escala deliberadamente seleccionada para equilibrar calidad de contenido, diversidad temática y viabilidad computacional en un entorno de entrenamiento con recursos limitados. Este rango es característico de proyectos de adaptación de dominio en entornos académicos o institucionales con recursos modestos.

Generación asistida del dataset

Dada la naturaleza intensiva en conocimiento de los dominios especializados, la construcción manual completa del dataset resultaba costosa en términos de tiempo y esfuerzo humano. Por ello, se adoptó un enfoque híbrido basado en generación asistida por modelos de lenguaje en línea, seguido de curación humana experta. Este enfoque metodológico es transferible a otros dominios especializados que compartan características de alta densidad conceptual. El proceso de generación asistida se estructuró en las siguientes etapas:

1. Fragmentación del corpus fuente: Los textos de referencia del dominio (documentación especializada, materiales técnicos y contenidos expertos) fueron segmentados en fragmentos manejables (chunks), considerando las limitaciones de ventana de contexto de los modelos generadores. Se utilizó una estrategia de segmentación semántica que preservaba la coherencia temática, con fragmentos típicamente de 1,000 a 2,000 tokens.

2. Diseño de instrucciones controladas: A los modelos generadores se les proporcionaron prompts explícitos que definían el rol esperado, el formato de salida y el nivel de concisión requerido. Los modelos empleados para la generación asistida incluyeron LLMs de frontera disponibles mediante APIs comerciales (específicamente Claude 3.5 Sonnet y GPT-4), seleccionados por su capacidad de seguir instrucciones complejas y mantener coherencia contextual.

3. Generación iterativa de pares Q&A: Para cada fragmento del corpus, se generaron múltiples pares de pregunta—respuesta, priorizando preguntas conceptuales y explicativas sobre preguntas triviales o puramente factuales. En promedio, se generaron entre 3 y 7 pares por fragmento de texto, dependiendo de su densidad informativa. El proceso completo de generación asistida produjo aproximadamente 21,000 pares candidatos antes de la curación humana.

Este enfoque permitió acelerar significativamente la fase inicial de creación del dataset, proporcionando una base estructurada sobre la cual aplicar validación humana. La tasa de generación fue aproximadamente 50-100 veces más rápida que la construcción manual completa, reduciendo el tiempo de construcción del dataset de varios meses a aproximadamente 2-3 semanas de trabajo efectivo.

Curación y revisión humana

A pesar de los avances en la generación automática de texto, los modelos de lenguaje pueden introducir errores factuales, interpretaciones incorrectas del contexto o imprecisiones conceptuales. Por este motivo, todos los pares Q&A generados fueron sometidos a revisión humana experta en el dominio. La curación se realizó mediante un proceso sistemático de 3 fases:

1. Un equipo de revisores con conocimiento del dominio realizó un primer filtrado.

2. Expertos en el dominio patrimonial revisaron los pares supervivientes.

3. Un subconjunto del 20% del dataset fue revisado independientemente por múltiples expertos para asegurar consistencia en los criterios de calidad.

El resultado final de este proceso de curación fue un dataset de entre 1,000 y 4,000 pares de alta calidad, representando una tasa de aceptación post-curación de aproximadamente 20-80% dependiendo de la rigurosidad aplicada y los objetivos específicos del proyecto. Este proceso garantizó que el dataset final mantuviera un estándar de calidad suficiente para el ajuste supervisado, al tiempo que preservó la riqueza contextual necesaria para el dominio especializado. Este procedimiento de validación humana es generalizable a otros dominios que requieran alta precisión conceptual, incluyendo campos legales, médicos, técnicos o académicos especializados.

Dataset como proxy de dominios especializados

El dataset patrimonial construido exhibe características representativas de dominios altamente especializados:

- Alta densidad de referencias contextuales implícitas
- Terminología técnica y específica del dominio
- Sensibilidad a errores de precisión conceptual
- Necesidad de coherencia semántica profunda
- Presencia de relaciones conceptuales complejas entre entidades

Estas propiedades hacen del patrimonio cultural un caso de estudio exigente, cuyos resultados son indicativos del comportamiento esperado en otros dominios especializados con características similares.

El dataset empleado presenta ciertas limitaciones inherentes a su diseño. En particular, su tamaño moderado y su validación predominantemente cualitativa limitan la posibilidad de realizar evaluaciones cuantitativas exhaustivas. No obstante, estas restricciones reflejan condiciones realistas de proyectos especializados en entornos académicos o institucionales con recursos limitados, y se alinean con el objetivo del estudio de evaluar la viabilidad técnica y metodológica del enfoque propuesto para la adaptación de dominio en general.

Los resultados obtenidos evidencian que el pipeline de supervised fine-tuning propuesto permite una adaptación efectiva del modelo Qwen 1.5 7B a dominios altamente especializados, incluso bajo restricciones computacionales significativas. Las mejoras observadas en el caso de estudio patrimonial son consistentes con el comportamiento esperado en otros dominios especializados con características similares de alta contextualidad y sensibilidad a errores de precisión. Una de las mejoras más notables observadas en el modelo ajustado fue su capacidad para incorporar contexto específico del dominio en sus respuestas. En comparación con el modelo base, el modelo afinado mostró una mayor precisión al abordar referencias especializadas y conceptos técnicos del dominio, reduciendo respuestas genéricas o descontextualizadas.

Tabla 3: Resultados cualitativos de la evaluación experta

Dimensión de Calidad	**Mejora (%)**	**Descripción**
Contextualización específica del dominio	70.8%	Capacidad de incorporar contexto especializado en las respuestas
Corrección factual y conceptual	53.6%	Reducción de errores factuales y conceptuales
Preferencia global del modelo ajustado	74%	Porcentaje de casos en los que el modelo ajustado fue considerado superior

El modelo base presentó, en múltiples casos, errores conceptuales atribuibles a generalizaciones excesivas o a la falta de conocimiento especializado del contexto del dominio. Tras el ajuste supervisado, estos errores se redujeron de manera apreciable, particularmente en preguntas que requerían un entendimiento implícito del contexto técnico o especializado. Desde una perspectiva de ingeniería, los resultados confirman que la combinación de técnicas PEFT, cuantización de baja precisión y optimizaciones de memoria permite llevar a cabo el fine-tuning de LLMs de 7B parámetros en hardware de consumo sin comprometer la estabilidad del entrenamiento.

Desde una perspectiva metodológica, las mejoras observadas son consistentes con las ventajas teóricas reportadas para las técnicas de ajuste eficiente en parámetros. Al actualizar únicamente un subconjunto reducido de parámetros mediante adaptaciones de bajo rango, el modelo puede internalizar patrones contextuales específicos del dominio sin alterar significativamente las capacidades lingüísticas generales adquiridas durante el pre-entrenamiento. La mayor mejora observada en la contextualización del dominio, en comparación con la corrección factual, puede explicarse por la naturaleza del dataset de entrenamiento utilizado. Los pares de pregunta–respuesta generados y curados se centran principalmente en explicaciones conceptuales y contextualización histórica del patrimonio cultural, más que en la simple recuperación de hechos aislados.

Como consecuencia, el modelo parece beneficiarse especialmente en su capacidad para situar las respuestas dentro de marcos contextuales culturalmente relevantes. Estos resultados son coherentes con investigaciones previas sobre adaptación de dominio en modelos de lenguaje, las cuales han demostrado que el supervised fine-tuning orientado a tareas específicas puede mejorar significativamente la alineación contextual del modelo incluso cuando se emplean datasets de tamaño moderado.

Limitaciones y trabajo futuro

A pesar de los resultados positivos, este estudio presenta varias limitaciones que deben ser reconocidas y que sugieren direcciones productivas para investigación futura.

Limitaciones metodológicas

Aunque la evaluación cualitativa sistemática empleada es particularmente adecuada para dominios especializados donde las métricas automáticas resultan insuficientes para capturar la calidad contextual profunda, limita la comparabilidad directa con otros trabajos que emplean evaluaciones cuantitativas extensivas. La implementación de métricas automáticas complementarias específicamente adaptadas a dominios especializados constituye una dirección valiosa para el trabajo futuro. La evaluación sistemática se realizó sobre 100 ejemplos. Si bien este tamaño es suficiente para identificar tendencias claras y fue validado por múltiples evaluadores expertos, un conjunto de evaluación más extenso proporcionaría mayor robustez estadística. El estudio se centró en validar la viabilidad del pipeline propuesto comparado con el modelo base, sin explorar comparaciones con enfoques alternativos de adaptación de dominio (como RAG, few-shot prompting con modelos más grandes, o diferentes arquitecturas PEFT). Estas comparaciones son valiosas, pero quedan fuera del alcance de este estudio metodológico.

Limitaciones del dataset

El dataset de 1,000-4,000 ejemplos, si bien suficiente para el objetivo del estudio, podría ampliarse para explorar una mayor diversidad temática y conceptual dentro del dominio especializado. Investigación futura podría explorar el comportamiento del enfoque con datasets de mayor escala. El dataset patrimonial se enfoca en ciertos aspectos del dominio. Una cobertura más exhaustiva de subdominios especializados (por ejemplo, conservación arquitectónica, patrimonio inmaterial, legislación patrimonial, etc.) podría mejorar la robustez del modelo. Aunque todos los ejemplos fueron validados por expertos humanos, el uso de LLMs para generación asistida puede introducir

sesgos sutiles o patrones específicos de los modelos generadores. La construcción de datasets mediante procesos alternativos (por ejemplo, extracción de diálogos reales de expertos) constituye una línea de investigación complementaria.

CONCLUSIONES

Este trabajo presentó un estudio de caso metodológico sobre el supervised fine-tuning eficiente de un modelo de lenguaje de 7B parámetros para tareas de pregunta–respuesta en dominios altamente especializados y contextualmente exigentes, demostrando que, mediante decisiones de ingeniería cuidadosamente fundamentadas, es posible adaptar LLMs de manera efectiva utilizando hardware de consumo. La integración de técnicas modernas de adaptación eficiente de parámetros y optimización computacional permitió entrenar adaptadores de reducido tamaño que produjeron mejoras claras y consistentes frente al modelo base, evidenciadas mediante una validación cualitativa experta sistemática que mostró incrementos sustanciales en contextualización específica del dominio, corrección factual y preferencia global. El caso de estudio en patrimonio cultural, seleccionado por su alta exigencia contextual y sensibilidad a errores conceptuales, confirmó la viabilidad técnica y metodológica del pipeline propuesto bajo restricciones computacionales realistas, al tiempo que sirvió como banco de pruebas representativo de otros dominios altamente especializados. Más allá de los resultados particulares, este trabajo aporta un marco de ingeniería completamente especificado y replicable que puede transferirse a contextos legales, médicos, técnicos o académicos con requisitos similares de precisión y contextualización. En conjunto, los hallazgos refuerzan la idea de que la especialización efectiva de modelos de lenguaje de gran tamaño no está limitada a infraestructuras de alto costo, sino que puede alcanzarse de manera sostenible mediante datasets de tamaño moderado y alta calidad combinados con técnicas eficientes de fine-tuning y validación cualitativa rigurosa.

REFERENCIAS

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Chen, T., Xu, B., Zhang, C., & Guestrin, C. (2016). Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and memory-efficient exact attention with IO-awareness. Advances in Neural Information Processing Systems, 35, 16344-16359.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient fine-tuning of quantized LLMs. Advances in Neural Information Processing Systems, 36.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8342-8360.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations.

Kestemont, M., Manjavacas, E., & Daelemans, W. (2022). LLMs in digital humanities research. Proceedings of the Computational Humanities Research Conference 2022, 311-326.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

Liu, S., Wang, C., Yin, H., Molchanov, P., Wang, Y. C., Cheng, K. T., & Chen, M. H. (2024). DoRA: Weight-decomposed low-rank adaptation. arXiv preprint arXiv:2402.09353.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1-67.

Schofield, A., Thompson, P., & Mimno, D. (2023). Humanities scholars and the use of large language models. Digital Scholarship in the Humanities, 38(4), 1567-1582.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.