ANTEPROYECTO DEL TRABAJO DE FIN DE GRADO

INFORMACIÓN GENERAL					
Alumno/a	Alejandro Silva Rodríguez				
Titulación:	Grado en Ingeniería de la Salud				
Tutor/es:	José Manuel Jerez Aragonés Francisco Javier Moreno Barea				
Título	Optimización de Modelos de Lenguaje Pequeños para Resumir Historias Clínicas: Comparativa, Destilación y RAG				
Subtítulo *(solo si en grupo)*					
Título en inglés	Optimizing Small Language Models for Clinical Summarization: Comparison, Distillation, and RAG				
Subtítulo en inglés *(solo si en grupo)*					
Trabajo en grupo:	**Sí**		**No**	X	
Otros integrantes del grupo:					

INTRODUCCIÓN

Contextualización del problema a resolver. Describir claramente de dónde surge la necesidad de este TFG y el dominio de aplicación. En caso de que el TFG se base en trabajos previos, debe aclararse cuáles son las aportaciones del TFG.

Las consultas clínicas e historiales reales suelen presentarse en forma de textos no estructurados, donde los profesionales de la salud frecuentemente copian y pegan información de consultas previas. Este método genera redundancia en los datos y dificulta la identificación de la información clínica actualizada y relevante, lo que puede impactar negativamente en la toma de decisiones médicas y en la eficiencia de los sistemas de gestión hospitalaria [1].

En los últimos años, los modelos de lenguaje de gran escala (LLMs, por sus siglas en inglés) han demostrado avances notables en la generación de resúmenes de alta calidad a partir de textos extensos [2]. Sin embargo, estos modelos suelen ser de gran tamaño y, en muchos casos, su uso está limitado por costos asociados a licencias propietarias. Como respuesta a esta limitación, se ha promovido el desarrollo de modelos de lenguaje pequeños (Small Language Models, SLMs) los cuales ofrecen accesibilidad y menores requerimientos computacionales. No obstante, estos modelos presentan desafíos significativos, como la omisión de información clave y entidades relevantes, especialmente cuando los documentos de entrada son extensos [3], [4].

Este trabajo surge de la necesidad de explorar soluciones eficientes para la generación de resúmenes clínicos que mantengan la relevancia de la información y sean accesibles en términos de costo y recursos computacionales. El dominio de aplicación de este estudio se encuentra en el procesamiento del lenguaje natural (NLP) aplicado al ámbito médico, con un enfoque en la mejora de los resúmenes clínicos generados por modelos de lenguaje de pequeños.

OBJETIVOS

Descripción detallada de en qué consistirá el TFG. En caso de que el objeto principal del TFG sea el desarrollo de software, además de los objetivos generales deben describirse sus funcionalidades a alto nivel.

Objetivo General
Desarrollar y evaluar modelos de lenguaje pequeños (Small Language Models, SLMs) en la generación de resúmenes clínicos, comparándolos con modelos de mayor escala como GPT-4, y optimizando su desempeño mediante técnicas avanzadas de procesamiento del lenguaje natural.

Objetivos Específicos

1. **Evaluación del desempeño de** SLMs
   - Analizar la capacidad de SLMs en la generación de resúmenes clínicos.
   - Comparar su rendimiento con modelos de mayor escala como GPT-4.
2. **Optimización de la generación de resúmenes**
   - Identificar técnicas que permitan mejorar la calidad de los resúmenes generados por SLMs.
   - Implementar estrategias de prompt engineering para optimizar la generación de resúmenes.
   - Aplicar técnicas avanzadas como RAG (Recuperador Generador), fine-tuning y distilación de conocimiento.
3. **Validación y evaluación de resultados**
   - Evaluar la calidad y relevancia de los resúmenes mediante la consulta con expertos médicos.
   - Aplicar métricas especializadas para medir la precisión y coherencia de los resúmenes:
     - **Métricas de superposición:** ROUGE, BLEU, METEOR.
     - **Métricas de similitud semántica:** BERTScore.
     - **Evaluaciones basadas en SLMs** para un análisis más preciso.
4. **Desarrollo de una aplicación prototipo**
   - Diseñar y desarrollar una aplicación web que implemente el mejor modelo identificado para la generación automática de resúmenes de historiales clínicos.

ENTREGABLES

Listado de resultados que generará el TFG (aplicaciones, estudios, manuales, etc.)

Informe comparativo del desempeño de los distintos modelos de lenguaje.

Modelo con mejor desempeño.

Prototipo aplicación web.

MÉTODOS Y FASES DE TRABAJO

METODOLOGÍA:

Descripción de la metodología empleada en el desarrollo del TFG. Especificar cómo se va a desarrollar. Concretar si se trata de alguna metodología existente y, en caso contrario, describir y justificar adecuadamente los métodos que se aplicarán.

Se adoptará una metodología iterativa basada en una comunicación continua con profesionales médicos. Este enfoque permitirá ajustar y refinar el desarrollo del proyecto en función del feedback recibido, asegurando que los resultados sean clínicamente relevantes. A través de ciclos de evaluación y mejora, se minimizarán los riesgos y se maximizará la utilidad práctica de los resúmenes generados.

FASES DE TRABAJO:

Enumeración y breve descripción de las fases de trabajo en las que consistirá el TFG.

1. **Revisión bibliográfica:** Se investigarán estudios previos sobre modelos de lenguaje, técnicas de resumen automático y métricas de evaluación para fundamentar el desarrollo del proyecto.
2. **Generación y validación de resúmenes modelo:** Se generarán resúmenes de referencia utilizando GPT-4 y se validará su calidad con un profesional médico para establecer un punto de comparación.
3. **Desarrollo y optimización de modelos SLMs:** Se aplicarán modelos de lenguaje pequeños (SLMs) y técnicas de prompt engineering para la síntesis de historiales clínicos. Además, se implementará la técnica RAG (Recuperador Generador) para mejorar la relevancia de los resúmenes y se realizará un proceso de destilación del mejor modelo.
4. **Evaluación comparativa:** Se compararán los resúmenes generados por SLMs con los de GPT-4. Para ello, se emplearán métricas especializadas como ROUGE, BLEU y BERTScore, asegurando una evaluación objetiva de la calidad del texto.
5. **Optimización final y refinamiento del modelo:** Se analizarán los resultados obtenidos en la fase comparativa y se aplicarán mejoras adicionales a los SLMs, optimizando su rendimiento antes de la implementación final.
6. **Desarrollo del prototipo de aplicación web:** Se diseñará e implementará una aplicación web en la que se integre el mejor modelo seleccionado para la generación automática de resúmenes clínicos, permitiendo su validación en un

entorno práctico.

7. **Comparación final y validación con expertos**: Se realizará una última evaluación de los resultados obtenidos, contrastándolos con la opinión de profesionales médicos para garantizar la utilidad y precisión del sistema.

TEMPORIZACIÓN:

La siguiente tabla deberá contener una fila por cada una de las fases enumeradas en la sección anterior. En caso de tratarse de un trabajo en grupo, se añadirá una columna HORAS por cada miembro del equipo. Debe especificarse claramente el número de horas dedicado por cada alumno/a y la suma de horas individual deberá ser también de 296.

FASE	HORAS *Alejandro Silva Rodríguez*
Revisión bibliográfica	**50**
Pruebas de ingeniería de prompts	**30**
Generación y validación de resúmenes modelo	**23**
Desarrollo y optimización de modelos SLMs	**50**
Evaluación comparativa	**20**
Refinamiento del modelo	**45**
Implementación de técnica RAG	**40**
Desarrollo del prototipo de aplicación web	**38**
	296

ENTORNO TECNOLÓGICO

TECNOLOGÍAS EMPLEADAS:

Enumeración de las tecnologías utilizadas (lenguajes de programación, frameworks, sistemas gestores de bases de datos, etc.) en el desarrollo del TFG.

Python
Pytorch
Hugging face
ollama
fastapi
javascript

RECURSOS SOFTWARE Y HARDWARE:

Listado de dispositivos (placas de desarrollo, microcontroladores, procesadores, sensores, robots, etc.) o software (IDE, editores, etc.) empleados en el desarrollo del TFG.

Visual studio code
Tarjeta de video Nvidia RTX 3050 laptop
Centro de Supercomputación y Bioinnovación (SCBI) de la Universidad de Málaga

REFERENCIAS

Listado de referencias (libros, páginas web, etc.)

[1] Searle, T., Ibrahim, Z., Teo, J., & Dobson, R. (2021). Estimating redundancy in clinical text. *Journal of Biomedical Informatics*, *124*, 103938. https://doi.org/10.1016/j.jbi.2021.103938

[2] Jin, H., Zhang, Y., Meng, D., Wang, J., & Tan, J. (2024). A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

[3] Zhang, G., Fukuyama, K., Kishimoto, K., & Kuroda, T. (2024). *Optimizing automatic summarization of long clinical records using dynamic context extension: Testing and evaluation of the NBCE method*. *arXiv preprint arXiv:2411.08586*.

[4] Grail, Q., Perez, J., & Gaussier, E. (2021). *Globalizing BERT-based transformer architectures for long document summarization*. En P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1792–1810). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.154

Málaga, 15 de Febrero de 2025

Firma tutor/tutora: Firma cotutor/a: Firma tutor/a coordinador/a: