Simplificación de textos médicos: Evaluación de legibilidad, relevancia y factualidad en PLS generados con Modelos de Lenguaje Abiertos Entrenados y Modelos Comerciales.

Desarrollo y despliegue de una herramienta web para profesionales de la salud.

Alberto Echeverria (a.echeverriaz@uniandes.edu.co)[a], Christian Palma (c.palma@uniandes.edu.co) [a], Paula Perdomo (p.perdomoe@uniandes.edu.co) [a], Gina Rios (g.riosr@uniandes.edu.co) [a]

[a] Estudiante de Maestría en Inteligencia Artificial, Universidad de los Andes, Bogotá, Colombia.

Resumen.

La alfabetización en salud es esencial para que los pacientes comprendan información médica y tomen decisiones informadas; sin embargo, casi la mitad de la población presenta dificultades para entender textos clínicos. Los Resúmenes en Lenguaje Sencillo (PLS) son una solución efectiva, pero su producción manual resulta lenta y costosa. Este trabajo aborda el desafío de automatizar la generación de PLS y, simultáneamente, desarrollar un sistema de evaluación que permita asegurar la calidad, factualidad y legibilidad de los resúmenes producidos por modelos generativos. El enfoque combina el análisis comparativo de cuatro modelos comerciales de última generación (GPT-5, Gemini 2.5 Flash/Pro y Claude Sonnet 4.5) y el ajuste fino de modelos abiertos pequeños (Llama-3.2-3B, Gemma-2-2B, Qwen-2.5-0.5B/2B) empleando tanto técnicas de entrenamiento supervisado (QLoRA) como de aprendizaje por refuerzo (DPO y PPO). Además, se entrenó un clasificador binario basado en DistilBERT para distinguir textos técnicos de textos simplificados, evaluando su capacidad de generalización en dominios externos.

Los resultados muestran que los modelos comerciales superan a los modelos abiertos en legibilidad y factualidad, aunque estos últimos alcanzan un rendimiento competitivo tras un ajuste cuidadoso que combina pérdidas CE+KL. Asimismo, el clasificador DistilBERT demostró la mejor capacidad de generalización. Como contribución final, se desarrolló una aplicación web que integra el modelo entrenado y permite generar y evaluar automáticamente la legibilidad de PLS, ofreciendo una herramienta práctica para mejorar la comprensión de textos médicos en entornos clínicos.

Palabras clave: LLM - Plain language summary – LoRA – Ajuste fino

1. Introducción

La alfabetización en salud se define como el grado en el que los individuos tienen la capacidad de obtener, procesar y entender la información y servicios básicos en salud, necesarios para tomar decisiones apropiadas (Ratzan et al., 2000). Según la Encuesta Europea de Alfabetización en Salud (Sørensen et al., 2015), el 47% de los encuestados, especialmente adultos mayores y personas con menor nivel educativo, presentan una alfabetización inadecuada.

En un entorno donde la participación del paciente y la transparencia de la información son cada vez más relevantes, es fundamental que la documentación sea clara y accesible (CDC, 2025). Aunque los resúmenes en términos sencillos (Plain Language Summaries, PLS) son una solución efectiva para traducir textos clínicos complejos a un lenguaje llano (Baedorf et al., 2020), su producción no automatizada es lenta, costosa y particularmente difícil en áreas con alta densidad terminológica.

Si bien los grandes modelos de lenguaje (LLM) tienen el potencial de automatizar la tarea de producir PLS, hasta ahora los estudios se han centrado en la generación de texto y han dejado en segundo plano la necesidad de desarrollar un sistema de evaluación de estos modelos. Por esto no se puede garantizar la calidad y fiabilidad de los resúmenes generados (Arias-Russi et al., 2025). Se buscó que el proyecto abordara esta brecha, generando resúmenes y desarrollando una métrica para evaluar los resúmenes generados. El impacto fue doble: a nivel técnico, se obtuvieron conclusiones sobre la efectividad del ajuste fino de modelos abiertos frente a modelos comerciales; a nivel social, se

espera que la herramienta mejore la alfabetización en salud, permitiendo que el personal de salud pueda simplificar textos técnicos para educar a sus pacientes.

2. Estado del arte

La simplificación de textos médicos ha evolucionado desde métodos manuales basados en reglas, centrados en la sustitución léxica y la reestructuración sintáctica, hasta técnicas más avanzadas que incorporan enfoques semánticos. Estudios pioneros (Ong et al., 2008; Kandula et al., 2010) aplicaron estas estrategias a historiales clínicos y materiales educativos, empleando recursos terminológicos como UMLS y MeSH para reemplazar términos complejos y dividiendo oraciones extensas o pasivas para mejorar la legibilidad. Aunque estas propuestas facilitaron la comprensión, seguían siendo transformaciones locales y dependientes de reglas predefinidas, sin capturar el significado global de los textos.

El salto cualitativo llegó con los modelos *transformer*, como BART y PEGASUS (Zhang et al., 2020), capaces de simplificar textos de forma generativa, aunque su aplicación biomédica se vio limitada por la falta de datos paralelos. Para afrontarlo, Devaraj et al. (2021) crearon el corpus Cochrane, que empareja resúmenes técnicos con versiones en lenguaje llano, lo que supuso un cambio conceptual al priorizar la comprensión de públicos no especializados. Posteriormente, Flores et al. (2023) mejoraron este enfoque incorporando métricas de legibilidad, consistencia factual y un *beam search* orientado a reducir alucinaciones.

Guo et al. (2022) crearon el corpus CELLS con resúmenes científicos y sus versiones en lenguaje llano, mientras que Attal et al. (2023) presentaron PLABA, con resúmenes biomédicos simplificados. En esta línea, Basu et al. (2023) desarrollaron Med-EASi con anotaciones detalladas. Lu et al. (2023) propusieron NapSS, un método en dos fases que primero resume el contenido con un modelo entrenado en pares de textos técnicos y sus versiones simplificadas, y luego utiliza las frases clave extraídas para generar prompts narrativos que guían al modelo durante la simplificación, manteniendo la coherencia discursiva. Además, dado que los corpus biomédicos disponibles siguen siendo de tamaño limitado, se han explorado estrategias para entornos de pocos datos, como métodos no supervisados (Laban et al., 2021) y el uso de aprendizaje por refuerzo. En esta línea, Rahman et al. (2024) desarrollaron el corpus *SimpleDC*, con textos sobre cáncer digestivo, que utilizaron para ajustar modelos LLaMA y mejorarlos mediante refuerzo guiado por un clasificador de lenguaje simple. Un trabajo reciente (Ferreira et al., 2025) propone un modelo capaz de "regular" la complejidad del texto generado.

Los modelos decoder-only, a diferencia de los encoder-decoder (ej. BART) adolecen de falta de alineación explícita entre entrada y salida (Kloser et al., 2024), lo que genera alucinaciones, omisiones, reescrituras excesivas y escaso control sobre términos clave y números, así como una alta dependencia de grandes cantidades de datos paralelos; para subsanar estas limitaciones se incorporan hoy mecanismos de control de contenido (copiado forzado de entidades, preservación de números y nombres), pérdidas de consistencia factual, etiquetas de legibilidad durante el entrenamiento, métodos de alineación implícita como aprendizaje contrastivo o prefix tuning, y evaluación humana continua para garantizar precisión y fidelidad. Estas mejoras son clave para que los *decoder-only* puedan alcanzar un nivel de precisión y control comparable al de los encoder–decoder, manteniendo sus ventajas en generación libre y aprovechando su escalabilidad en escenarios con pocos datos.

3. Metodología

La metodología usada fue Crisp-DM clásica para el desarrollo y la propuesta de solución, que consistió en los siguientes pasos: preparación de los datos, comparación de métricas de cuatro modelos comerciales con textos médicos de Cochrane, fine-tuning de cuatro modelos pequeños abiertos y evaluación de desempeño para generación de PLS, entrenamiento de un clasificador binario que diferencia un texto simplificado de uno técnico, y el despliegue con un API pensada para ser usada por personal de salud.

3.1 Preparación de los datos

Los datos utilizados en este proyecto fueron suministrados por la Universidad de los Andes y fueron publicados originalmente en la investigación "Bridging the Gap in Health Literacy: Harnessing the Power of Large Language Models to Generate Plain Language Summaries from Biomedical Texts" (Arias-Russi et al., 2025). La base de datos inicial constaba de 14441 textos, en formato .txt, tanto en lenguaje simplificado (PLS), como en lenguaje médico (No PLS), recolectados de Clinicaltrials.gov, la librería Cochrane, la página pública de estudios de Pfizer y de Citeline Regulatory. El data set total tenía 61354 textos, dividido en grupo de entrenamiento y grupo de prueba. Al realizar la exploración de los datos se evidenció que solo los textos de Cochrane tenían parejas No PLS-PLS, por lo tanto, solo se utilizaron esos datos para el entrenamiento.

Para la tarea de simplificación de texto médico, se realizó una normalización leve del texto con regex, dejando símbolos, puntuaciones, acrónimos y abreviaciones, por su significado lingüístico y médico. Posteriormente, se protegieron las palabras en mayúsculas, se normalizaron los espacios en blanco y se tokenizó el texto. Para la tarea de clasificación, el balance de clases es vital, y en nuestro caso, al escoger los datos de Cochrane como única fuente, obtuvimos un data set balanceado. Al finalizar la exploración y el preprocesamiento de los datos, se crearon dos DataFrames de parejas No PLS-PLS, uno en texto normalizado y otro en texto tokenizado. Estos DataFrames, junto con los datos originales, fueron depositados en un bucket en S3, desde donde se accedieron para los experimentos.

3.2 Modelos comerciales

Siguiendo la metodología experimental descrita por Arias-Russi et al. (2025), se generaron Resúmenes en Lenguaje Sencillo (PLS) a partir de un corpus validado de 300 resúmenes técnicos de Cochrane, utilizando el prompt estandarizado del estudio original (Anexo 1) en cuatro modelos de lenguaje de última generación: GPT-5, Gemini 2.5 (versiones Flash y Pro) y Claude Sonnet 4.5.

Para cada modelo se calculó el promedio de diferentes scores. Los scores de legibilidad que se computaron, usando la biblioteca de Python Readability, fueron: Flesch Kincaid grade, Coleman Liau index, Flesch Reading ease, Gunning Fog index, SMOG index y Dale Chall score. Además, se calculó AlignScore (Zha et al., 2023) como métrica de factualidad y BERTScore (Zhang et al., 2020) para relevancia.

3.3 Fine-tuning de modelos pequeños abiertos

La metodología empleada para el ajuste fino de los modelos abiertos se basó en un análisis previo de las diferentes variantes existentes para poder especializar el modelo a la tarea planteada de simplificación de textos. Esta tarea requiere un nivel de abstracción alto, a diferencia de la tarea convencional de simplificación de texto, ya que la información correspondiente a cada una de las cuatro secciones en los que se estructura el PLS, no está en el mismo orden en el abstract de origen. Las pruebas iniciales se realizaron con el modelo Llama 3B-Base (Llama3.2 3B), y los resultados “zero shot” del mismo, con el prompt C1 (Anexo 1) (Arias-Russi et al., 2025).

Posteriormente, se optó por utilizar el modelo Instruct, el cual tiene un alineamiento inicial, adecuándolo para la tarea de abstracción y transformación en un texto simplificado y bien estructurado. Se buscó la forma de mejorar este modelo: induciendo al modelo a un estilo “más humano”, cercano a los PLS de referencia, y adecuando el estilo al grado de legibilidad y complejidad exigido.

Se consideraron dos alternativas de entrenamiento, no excluyentes: un ajuste fino mediante QLoRA, que denominaremos SFT (Supervised Fine Tuning), y un ajuste mediante aprendizaje por refuerzo o RLHF (Reinforcement Learning with Human Feedback).

Para el entrenamiento SFT, se tomaron 2000 parejas No PLS-PLS de Cochrane. Con el objeto de maximizar los medios disponibles, se limitó en los prompts el número de tokens de entrada, eliminando contenido de la parte central (priorización *head and tail*), manteniendo al máximo la comprensión de la entrada.

Durante el SFT, se consideraron las siguientes variables en la función de costo compuesta: un término de coseno que asegura alinear semánticamente el PLS con el abstract, un término de entropía cruzada (Cross Entropy, CE) que asegura un aprendizaje por autorregresión y, un término de divergencia KL (Kullback–Leibler) que asegura que el modelo resultante no se aleje del modelo de referencia, el Llama 3.2B Instruct.

El término de CE empleado, para el caso en el que solo se consideró este término, se implementó con *label smoothing* y filtrando únicamente la salida de generación con el label -100, así:

$$\mathcal{L}_{\mathcal{CE}} = -\sum_{i=1}^{K} \left[(1-\epsilon)\, y_i + \frac{\epsilon}{K} \right] \log p_i$$

Donde ϵ es el parámetro de suavizado que redistribuye parte de la probabilidad de la clase correcta hacia el resto del vocabulario, K es el número total de clases (tamaño del vocabulario), y_i es el valor de la etiqueta *one-hot* para la clase i (1 si es la clase correcta, 0 si no), y p_i es la probabilidad estimada por el modelo para la clase $i$después de aplicar softmax.

El término de divergencia KL se complementó al término de CE, pero en este caso no se hizo *label smoothing*. siendo la función de costo:

$$\mathcal{L} = \mathcal{L}_{\mathcal{CE}} + \lambda_{KL}\, \mathcal{L}_{\mathcal{KL}}$$

$$\mathcal{L} = -\sum_{t=1}^{T} \log p_\theta\,(\, x_t \mid x_{<t}\,) + \lambda_{KL}\, \tau^2 \sum_{i=1}^{K} \left(q_\tau(\, i \mid x\,)\, \log ! \frac{q_\tau(\, i \mid x\,)}{p_\tau(\, i \mid x\,)} \right)$$

Se tomó un valor de λ_{KL} de 0.1. Para el SFT, se estableció un análisis de parámetros en los que se incluyó el término de tasa de aprendizaje ($4e^{-5}$ y $5e^{-6}$) y, opcionalmente, el término de KL. Se empleó una temperatura fija $\tau = 2$ en el término KL de destilación (Hinton et al., 2015).

Para el fine-tuning, se empleó un esquema QLoRA, cargando el modelo base cuantizado a 4 bits (nf4) con *double quantization* para reducir memoria. Sobre este backbone se aplicaron adaptadores LoRA (r = 16, α = 32, dropout = 0.1) inyectados en los módulos q/k/v/o-proj del transformador, permitiendo entrenar únicamente un bajo número de parámetros eficaces. El modelo fue preparado con prepare_model_for_kbit_training, manteniendo el resto de los pesos congelados, y el entrenamiento se realizó en configuración CAUSAL LM, con *tokenizer* alineado y pad_token = eos_token.

Para el entrenamiento con RLHF, se realizaron pruebas mediante la técnica de DPO (Direct Preference Optimization), técnica híbrida entre SFT y RL, y con la técnica clásica de PPO (Proximal Policy Optimization), que permite ir aprendiendo en función de las recompensas y utiliza la función de valor.

Para el caso de DPO (Rafailov et al., 2023) se generaron parejas de PLS, combinando el PLS de referencia de Cochrane y el PLS generado con el modelo resultante de SFT, alimentado por el prompt C1 completo. Las primeras se etiquetan con *chosen* y las segundas con *rejected*, y de esta manera el entrenamiento fue asignando recompensas. Se ajustó directamente la probabilidad relativa entre ambas respuestas, incrementando la preferencia del modelo por las alternativas etiquetadas como *chosen* sin necesidad de estimar una función de recompensa explícita.

Para el caso de PPO, los datos se alimentaron invocando al prompt para la generación de PLS nuevos, que fueron evaluados, generando la siguiente métrica:

$$Reward = 0.5\, AlignScore + 0.3\, BERTScore_{F1} + 0.2\, Readability$$

3.4 Clasificador binario

Para el proceso de clasificación binaria de textos médicos se entrenaron tres modelos: uno con arquitectura de red neuronal basada en DistilBERT (Sanh et al., 2019), la cual utiliza embeddings contextuales; y dos que utilizan embeddings dispersos: una regresión logística y un Random Forest. El modelo DistilBERT usado fue 'distilbert-base-uncased', el cual es una versión destilada del modelo BERT base, siendo más pequeño, más rápido y eficiente que el modelo original. Para la tarea de clasificación se utilizó un cabezal de clasificación con dos capas densas que actúan como una red adicional.

Para el entrenamiento de los 3 modelos se integraron 2 fuentes de datos: Cochrane y PLOS, siendo Cochrane el núcleo del entrenamiento. Se utilizaron 3426 pares válidos de textos PLS y no PLS, de los cuales se dejaron 5481 registros para entrenamiento y 1371 para validación. Se enriqueció el entrenamiento con 5000 textos simplificados de PLOS y 5000 técnicos.

Para los modelos con embeddings dispersos, se realizó tokenización con Tf-idf con 15000 parámetros y n-gramas con rango de 1 a 3. Para el entrenamiento de la regresión lineal se utilizó un parámetro de regularización C = 1 y solver liblinear; el del Random Forest se realizó con 100 estimadores y con balance de clases. El entrenamiento del modelo DistilBERT consistió en un fine-tuning suave, basado en AdamW y un learning rate base de 2e-5. Se realizó un entrenamiento por 5 épocas en donde se garantizó convergencia analizando las métricas de rendimiento que iba produciendo el optimizador.

Para verificar y garantizar la eficacia del modelo en contextos distintos, se realizó una prueba con textos simplificados y técnicos de eLife. Esta es una excelente manera de verificar si aspectos más complejos de la tarea de clasificación como densidad de jerga y marcadores de discurso han sido reconocidos por DistilBERT, mostrando su capacidad de generalización.

3.5 Despliegue de la aplicación

Como fase de integración tecnológica, se llevó a cabo el desarrollo y despliegue de una aplicación web diseñada para facilitar la interacción de los profesionales de la salud con el modelo generativo. La arquitectura de la solución (Figura 1) se estructuró en dos componentes principales: un backend desarrollado en Python mediante el framework FastAPI y alojado en una instancia de computación en la nube AWS EC2, el cual gestiona las peticiones de inferencia utilizando el modelo Llama-3.2-3B-Instruct previamente entrenado y cuyos pesos se almacenaron en un repositorio de objetos AWS S3. Complementariamente, se implementó un frontend estático basado en tecnologías estándar (HTML, CSS y JavaScript), hospedado también en AWS S3, garantizando así una interfaz accesible y escalable. El código fuente de la implementación se encuentra disponible en el repositorio del proyecto (https://github.com/paula-perdomo/simplificar-textos-medicos/tree/main/app). La herramienta permite a profesionales de la salud generar un PLS en base a un texto médico. Además, provee un cálculo de métricas de legibilidad sobre ambos textos que permite validar si el PLS generado por el modelo entrenado es de fácil comprensión.

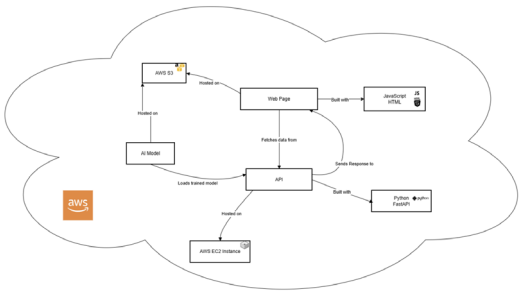

Figura 1. Esquema de la arquitectura de la solución. Imagen creada por los autores.

4. Resultados

4.1 Modelos comerciales

La evaluación cuantitativa (Tabla 1) muestra principalmente el trade-off entre legibilidad y consistencia factual.

Tabla 1. Evaluación de modelos comerciales.

Modelo	Legibilidad						Factualidad	Relevancia
	CLI↓	FRE↑	GFI↓	SMOG↓	FKGL↓	DCRS↓	AlignScore↑	BERTScore↑
gpt-5	5.84	94.70	5.37	6.33	2.33	4.07	0.5501	0.8327
gemini-2.5-flash	7.24	85.90	6.22	7.12	3.53	4.20	0.6037	0.8330
gemini-2.5-pro	8.14	82.59	7.71	8.11	4.61	4.27	0.6386	0.8298
claude-sonnet-4-5-20250929	10.1	65.75	8.09	9.46	6.66	8.14	0.4787	0.8448

GPT-5 demostró un desempeño superior en todas las métricas de legibilidad, logrando generar textos con una complejidad bastante baja (Flesch Reading Ease de 94.70 y Flesch-Kincaid Grade Level de 2.33). Sin embargo, en términos de consistencia factual, Gemini 2.5 Pro obtuvo el mejor rendimiento (AlignScore de 0.6386), seguido de cerca por su versión Flash. Por otro lado, Claude Sonnet 4.5 obtuvo mejor rendimiento en relevancia semántica (BERTScore de 0.8448).

4.2 Modelos abiertos

Las pruebas iniciales realizadas con el modelo Llama 3B-Base (Llama3.2 3B), no dieron resultados aceptables, en cuanto a capacidad de abstracción y orden de las secciones. Por esto, se optó por utilizar el modelo Instruct, notándose una marcada mejoría del desempeño.

Durante el entrenamiento SFT, al crear la función de costo compuesta, los términos basados en similitud coseno no produjeron la alineación prevista entre *hidden states* y resúmenes abstractos. La proyección 2048→384 supone una compresión agresiva que elimina componentes semánticos finos necesarios para la separación en el espacio latente. Este módulo operó como un cuello de botella, al introducir una transformación no co-adaptada con el backbone principal y limitar la transferencia de gradiente útil. Su ubicación, externa a la arquitectura principal, provocó un desacoplamiento estructural, de modo que el modelo a entrenar no ajustó sus representaciones internas al objetivo coseno, reduciendo así la separabilidad necesaria para obtener mejoras en AlignScore.

Tabla 2. Resultados de las diferentes estrategias de entrenamiento del modelo Llama 3.2 3B.

						Legibilidad		Complejidad
	Parámetros	MetaScore	PPL	BERT	ALIGN	FKGL	SMOG	DCRS
PLS humano		**0,753**	**14,787**					
Llama-3.2-3B	*Ref (Arias-Russi et al.)*			*0,849*	*0,878*	*15,73*	*10,79*	*9,39*
Llama-3.2-3B Instruct	BASELINE	0,737	10,328	0,852	0,884	12,445	13,89	12,029
Instruct								
Yes	CE 1 LRHi	0,7318	9,2683	0,857	0,855	12,152	13,421	11,615
Yes	CE 2 LRHi	0,7112	9,71624	0,855	0,895	12,05	13,791	11,57
Yes	CE 3 LRHi	0,7200	10,230	0,857	0,903	12,525	14,26	11,683

Yes	CE 4 LRHi	0,7293	121	0,856	0,884	12,621	14,058	11,714
Yes	CE 2 LRLow	0,739	10,378	0,849	0,872	11,379	13,04	11.219
Yes	CE 4 LRLow	0,733	8,649	0,849	0,904	12,901	14.24	12,212
Yes	CE KL 2 LRLow	0,734	9,920	0,849	0,887	12,178	13,64	11,675
Yes	**CE KL 4 LRLow**	**0,737**	**10,818**	**0,849**	**0,872**	**11,266**	**12,97**	**11.143**
Yes	CE KL 6 LRLow	0,7343	10,006	0,849	0,876	11,765	13,329	11,293
Yes	CE KL 10 LRLow	0,7393	10,3937	0,85	0,878	11,268	12,95	11,211
No	CE KL 6 LRLow	0,7425	9,0764	0,851	0,862	10,708	12,716	10,954

La estrategia de entrenamiento seleccionada por su mejor desempeño fue aquella con la función de pérdida compuesta por CE+KL y una tasa de aprendizaje baja (Tabla 2). Estos resultados son los que muestran un equilibrio óptimo entre factualidad, relevancia, complejidad y legibilidad. Los resultados no llegan a degenerarse (generaciones incorrectas) como ocurre con los modelos con altas épocas y función de costo basada únicamente en Cross Entropy (High CE 4epoch). Esta estrategia de entrenamiento permite afinar el estilo original (humano) de los resúmenes sin perder el orden de secciones y capacidad de abstracción y orden del modelo Instruct.

Como se observa en la curva de aprendizaje (Figura 2), ésta va descendiendo de forma gradual y estable, gracias al efecto de regularización que ejerce el modelo *teacher Instruct*, en comparación con un modelo que únicamente considera *Cross Entropy*.

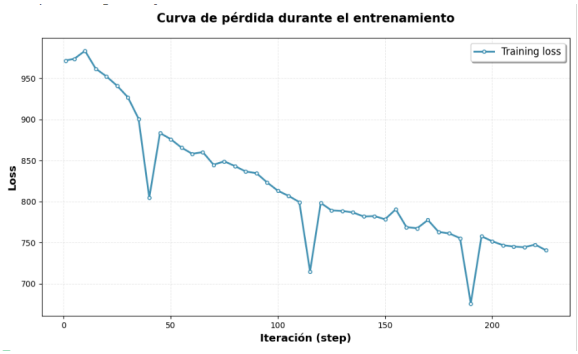

Figura 2. Curva de aprendizaje del Caso de CE+KL (6 epoch) y LR Low. Imagen creada por los autores.

La media de los valores de Alignscore y BERTScore se mantiene alta (mayor a 0.8). Este resultado, muestra que el PLS generado tiene buena factualidad (Alignscore) y relevancia (BERTscore), como resultado del buen comportamiento del modelo Instruct. Además, se observó una ligera mejoría en los valores de legibilidad. Conviene señalar, que en la tabla no es fácil apreciar un modelo que destaque. Por ello, se optó por analizar ejemplos para evaluar la calidad de estos. Este requeriría un gran esfuerzo de etiquetado, con el que actualmente no contamos, por lo que se creó una estrategia para evaluar de forma automática la calidad de los PLS, que llamamos MetaScore (Anexo 2), basándonos en diferentes propuestas encontradas en la literatura, tales como "*LLM as a judge*" (Zheng et al., 2023) o el propuesto por Gemini (Guidroz et al., 2025). De igual forma, se utilizó la métrica de perplejidad para evaluación de la generación. La fórmula del MetaScore, inspirado en Liu et al. (2023), es:

$$MetaScore = f(BERT\_fact, AlignScore, Readability, Coherence, Conciseness, PPL))$$

Con el objeto de analizar la calidad del resumen generado, observamos que el modelo Instruct base presenta valores muy cercanos a la media de los PLS de referencia, y un nivel de perplejidad superior a 10. A medida que se hace el ajuste fino, éste produce una disminución de la perplejidad, lo que resulta en un texto más fluido que el producido por

el Instruct base. Hay que señalar que el modelo solo con CE e Instruct base con 4 epochs y tasa de aprendizaje alta presenta cierta degeneración en la generación de texto (perplejidad de 121).

Usando la estrategia de entrenamiento escogida se entrenaron otros 3 modelos con prestaciones parecidas a Llama3.2 3B Instruct y se comparó el desempeño comparativo de los mismos (Tabla 3). Analizando los resultados se escogieron dos modelos para el despliegue: el Llama 3.2-3B entrenado con CE+KL y 4 Epoch con una tasa de aprendizaje de $5e^{-6}$, que da un buen equilibrio de MetaScore y perplejidad, con valores mejorados de legibilidad, y complejidad respecto al Baseline de Llama3.2 3B-Instruct "zero-shot" y el PLS humano. A pesar de que las métricas del modelo entrenado con CE+KL y 6 Epochs fueron mejores, el tiempo de inferencia en el despliegue era mucho mayor y la estructura del texto generado no era adecuada el 90% de las veces; el modelo Gemma 2-2B Instruct por su rendimiento excepcional en cuanto a legibilidad.

Tabla 3. Evaluación de los modelos abiertos.

Modelo	**Legibilidad**						**Factualidad**	**Relevancia**
	CLI↓	FRE↑	GFI↓	SMOG↓	FKGL↓	DCRS↓	AlignScore↑	BERTScore↑
Gemma-2-2B-Instruct	11.887	54.979	11.128	11.428	9.309	10.263	0.819	0.838
Qwen-2.5-0.5B-Instruct	15.569	21.615	15.547	13.603	13.964	11.044	0.6037	0.814
Llama-3.2-3B-Instruct	13.76	43.39	13.453	12.97	11.26	11.14	0.8785	0.849
Qwen-2.5-2B-Instruct	14.923	33.669	14.985	13.908	12.495	11.589	0.816	0.83

4.2.1 Reinforcement Learning with Human Feedback

Para el entrenamiento con RLHF con la técnica DPO no se logró convergencia porque las recompensas de "*chosen*" no mostraron una tendencia creciente estable, sino oscilaciones sin mejora sostenida. Al mismo tiempo, las recompensas de "*rejected*" se mantuvieron relativamente altas, lo que indicó que el modelo no estaba aprendiendo a separar claramente buenas de malas respuestas. La distancia *reward(chosen) − reward(rejected*) no aumentó, así que el optimizador no recibió gradientes consistentes. En conjunto, esto sugiere señal de preferencia débil o ruido en el modelo/reward, impidiendo que DPO encontrara una política dominante.

Al entrenar con técnica PPO, se indujo olvido catastrófico (*catastrophic forgetting*), con la red de políticas colapsando hasta generar caracteres Unicode aleatorios. Esto se debió probablemente a restricciones insuficientes de divergencia KL durante el aprendizaje por refuerzo, permitiendo que las actualizaciones de política se alejaran demasiado de la inicialización del fine-tuning supervisado.

4.3 Clasificador binario

El rendimiento de los clasificadores binarios fue evaluado con la métrica precisión o accuracy (Tabla 4). La evaluación se realizó en los datos de validación de Cochrane con un proceso de maximización del umbral de clasificación de 0.98. Estos procesos de recalibración en cambios de dominios están bien documentados en Kamath et al. (2025).

La precisión del modelo de regresión logística fue de 1 y su desempeño en contextos distintos alcanzó una precisión de 83.2%. La precisión del Random Forest fue 0.99 con una capacidad de generalización más pobre, evidenciada por la precisión de 58.71% en contextos distintos.

La precisión del modelo DistilBERT es de 0.99 en la validación con el data set de Cochrane. Al verificar el desempeño en contextos distintos, la precisión del modelo fue de 87.14%.

Tabla 4. Desempeño de Clasificadores Binarios.

Modelo	Precisión (Cochrane)	Precisión (eLife)
Regresión logística	1.00	83.20%
Random Forest	0.99	58.71%
DistilBERT	0.99	87.14%

5. Discusión

Al comparar el desempeño de los modelos comerciales (API-based) entre sí, es evidente que el modelo GPT-5 es el que produce los resúmenes más legibles mientras que sus resultados de factualidad y relevancia son bajos. Inversamente, el modelo Claude Sonnet tiene la más baja legibilidad, pero la mayor relevancia, sugiriendo que mientras algunos modelos priorizan la simplificación sintáctica extrema, otros mantienen una mayor alineación con los hechos técnicos del texto fuente.

Los resultados de legibilidad sobrepasan los obtenidos por Arias-Russi et al. (2025), siendo los valores de relevancia muy similares. Esta mejora se da ya que los modelos utilizados en el presente trabajo son más grandes, con más parámetros de entrenamiento. Sin embargo, no se logró igualar ni mejorar el AlignScore con ninguno de estos modelos. Las posibles causas son: diferencias en el preprocesamiento de los datos, la tokenización, el tamaño de los batches o incluso por diferencias sintácticas como la superposición semántica.

Al comparar el desempeño de los modelos abiertos entre sí, se evidencia un marcado trade-off entre la capacidad de simplificación y la fidelidad del contenido. El modelo Gemma-2-2b-Instruct demostró un desempeño superior en las métricas de legibilidad, logrando reducir significativamente la complejidad del texto (Flesch Reading Ease de 54.979 y Flesch-Kincaid Grade Level de 9.309), lo que lo posiciona como el más apto para generar textos accesibles. Por el contrario, Llama-3.2-3B-Instruct obtuvo el mejor rendimiento en términos de consistencia factual y relevancia semántica, alcanzando los valores más altos en AlignScore (0.8785) y BERTScore (0.849), aunque a costa de una mayor complejidad sintáctica. En cuanto a la familia Qwen, se observa que el tamaño de los parámetros influye drásticamente en la calidad: mientras que Qwen-2.5-2B-Instruct mantiene un equilibrio aceptable con una factualidad competitiva (AlignScore de 0.816), el modelo más pequeño, Qwen-2.5-0.5B-Instruct, sufre una degradación crítica en la consistencia factual (AlignScore de 0.6037), sugiriendo que, a pesar de su eficiencia, carece de la capacidad necesaria para retener la información técnica precisa requerida en el dominio biomédico.

Si se compara el desempeño de los modelos comerciales con los modelos abiertos, se observa la superioridad de los primeros en la generación de texto, tanto en legibilidad global como en factualidad y relevancia. Los valores de FKGL y SMOG son sistemáticamente más bajos, lo que indica que los modelos comerciales producen texto más simple y fácil de entender. De igual forma, el AlignScore es más alto y el BERTScore es similar, lo que nos indica que el texto generado por los modelos comerciales tiene más coherencia semántica.

El modelo DistilBERT fue el escogido para la tarea de clasificación para el despliegue ya que, a pesar de que el desempeño de los tres modelos es muy similar al evaluarse con los datos de Cochrane, al comparar la capacidad de generalización con los datos de eLife, se observa que el modelo DistilBERT es superior, y en contextos de producción se desea mayor capacidad de generalización. Además, su tiempo de inferencia es muy corto (120ms). Hay otros tres factores que hacen que este modelo sea ideal para el despliegue:

- Retención de rendimiento: DistilBERT retiene aproximadamente el mismo rendimiento en comparación a modelos específicamente entrenados en textos médicos como ClinicalBERT, específicamente en tareas de clasificación, pero con menos capas y parámetros. (Karim et al., 2024)
- El tiempo de inferencia es un 60% más rápido en comparación a BERT (Adel et al., 2022), esto permite que el modelo pueda ser levantado en contextos de producción con tiempos bajos de inferencia, como es el caso del producto final del proyecto.
- Semántica requerida: Entender la diferencia entre un lenguaje técnico y un lenguaje plano recae en reconocimientos de longitud de oraciones, densidad de jerga y marcadores de discurso. Estas tareas no necesitan de las 12 capas que BERT posee, con las 6 capas que posee DistilBERT (Figura 3) se llegan a resultados similares.

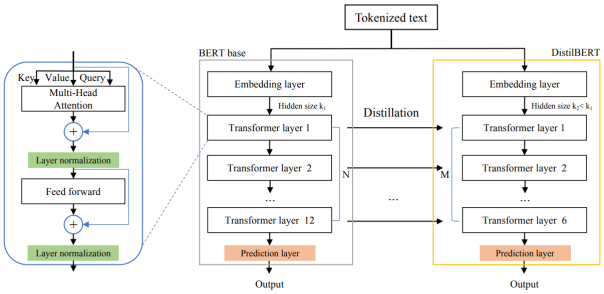

Figura 3. Componentes y arquitectura del modelo DistilBERT. Imagen de Adel et al. (2022).

El desempeño del modelo en el data set de Cochrane alcanza y supera valores encontrados en clasificadores similares en la literatura como el presentado en 2023 por Mijatović et al., donde lograron una precisión de 0.87 en el set de validación balanceado; o el de Arias-Russi et al. (2025), donde lograron una precisión del 0.97. En comparación con este último trabajo, no logró igualar la precisión de generalización (0.95 eLife+PLOS vs. 0.87 eLife).

6. Conclusiones y trabajo futuro

En conclusión, los pequeños modelos de lenguaje abiertos tienen un desempeño satisfactorio al compararse con los modelos comerciales grandes, confirmando que no se requieren grandes cantidades de parámetros para entrenar un modelo y crear una solución a un problema. Nuestro modelo, aunque aún tiene espacio para mejoras, puede ser utilizado de manera satisfactoria en el ambiente médico para continuar el esfuerzo de la alfabetización en salud.

Dentro de las oportunidades de futuras líneas de investigación, se propone la estandarización del MetaScore para su uso en generación de texto. Además, se puede profundizar en el desarrollo del modelo abierto pequeño, mejorando el prompt, entrenándolo con textos diferentes a Cochrane y puliendo la generación con técnicas de aprendizaje por refuerzo.

Como limitaciones de nuestro trabajo, nos encontramos con que, por las características del proyecto, y los tiempos predeterminados, hacer un análisis cualitativo de los textos generados por los modelos abiertos no fue posible. Además, por nuestra capacidad de hardware, el tiempo de inferencia de la API sigue siendo elevado.

Referencias

Adel, H., Dahou, A., Mabrouk, A., Abd Elaziz, M., Kayed, M., El-Henawy, I. M., Alshathri, S., & Amin Ali, A. (2022). Improving Crisis Events Detection Using DistilBERT with Hunger Games Search Algorithm. *Mathematics*, *10*(3), 447. https://doi.org/10.3390/math10030447

Arias-Russi, A., Salazar-Lara, C., & Manrique, R. (2025). Bridging the Gap in Health Literacy: Harnessing the Power of Large Language Models to Generate Plain Language Summaries from Biomedical Texts. 269–284. https://doi.org/10.18653/v1/2025.cl4health-1.23

Attal, K., Ondov, B., & Demner-Fushman, D. (2023). A dataset for plain language adaptation of biomedical abstracts. *Scientific Data, 10*(1). https://doi.org/10.1038/s41597-022-01920-3

Baedorf, S., Bahador, B., Gawrylewski, H., & Gertel, A. (2020). Promoting equity in understanding: A cross-organizational plain language glossary for clinical research [Review of Promoting equity in understanding: A cross-organizational plain language glossary for clinical research]. *Medical Writing*, 29(4), 10–15.

Basu, C., Vasu, R., Yasunaga, M., & Yang, Q. (2023). Med-EASi: Finely Annotated Dataset and Models for Controllable Simplification of Medical Texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(12), 14093–14101. https://doi.org/10.1609/aaai.v37i12.26649

CDC. (2025, July 25). Plain Language Materials & Resources. Health Literacy. https://www.cdc.gov/health-literacy/php/develop-materials/plain-language.html

Devaraj, A., Marshall, I. J., Wallace, B. C., & Li, J. J. (2021). Paragraph-level simplification of medical texts. *NAACL*, 4972–4984. https://doi.org/10.18653/V1/2021.NAACL-MAIN.395

Ferreira, D. J. B., Almeida, T. M., & Matos, S. (2025). A framework for fine-grained complexity control in health answer generation. *In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 1111–1131)*. Association for Computational Linguistics. https://aclanthology.org/2025.acl-srw.87.pdf

Flores, L. J., Huang, H., Shi, K., Chheang, S., & Cohan, A. (2023). Medical text simplification: Optimizing for readability with unlikelihood training and reranked beam search decoding. *Findings of ACL: EMNLP* 2023, 4859–4873. https://doi.org/10.48550/arXiv.2310.11191

Guidroz, T., Ardila, D., Li, J., Mansour, A., Jhun, P., Gonzalez, N., Ji, X., Sanchez, M., Kakarmath, S., Bellaiche, M. M. J., Garrido, M. Á., Ahmed, F., Choudhary, D., Hartford, J., Xu, C., Serrano Echeverria, H. J., Wang, Y., Shaffer, J., Cao, E., Matias, Y., Hassidim, A., Webster, D. R., Liu, Y., Fujiwara, S., Bui, P., Duong, Q. (2025). LLM-based Text Simplification and its Effect on User Comprehension and Cognitive Load. arXiv preprint arXiv:2505.01980.

Guo, Y., Qiu, W., Leroy, G., Wang, S., & Cohen, T. (2022). CELLS: A Parallel Corpus for Biomedical Lay Language Generation. *ArXiv.org*. https://arxiv.org/abs/2211.03818v1

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531.*

Kandula, S., Curtis, D., & Zeng-Treitler, Q. (2010). A semantic and syntactic text simplification tool for health content. *AMIA Annual Symposium proceedings. AMIA Symposium*, *2010*, 366–370.

Karim, A. A. J., Asad, K. H. M., & Alam, M. G. R. (2024). Larger models yield better results? Streamlined severity classification of ADHD-related concerns using BERT-based knowledge distillation. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0315829

Klöser, L., Beele, M., Schagen, J.-N., & Kraft, B. (2024). German text simplification: Finetuning large language models with semi-synthetic data. I*n Proceedings of the Fourth Workshop on Language Technology for Equality,*

Diversity, Inclusion (pp. 63–72). Association for Computational Linguistics. https://aclanthology.org/2024.ltedi-1.7/

Laban, P., Schnabel, T., Bennett, P., & Hearst, M. A. (2021). Keep It Simple: Unsupervised Simplification of Multi-Paragraph Text. *ArXiv (Cornell University)*, 6365–6378. https://doi.org/10.18653/v1/2021.acl-long.498

Liu, Y., Wang, S., & He, P. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv preprint arXiv:2303.16634.*

Lu, J., Li, J., Wallace, B., He, Y., & Pergola, G. (2023). NapSS: Paragraph-level Medical Text Simplification via Narrative Prompting and Sentence-matching Summarization. https://doi.org/10.18653/v1/2023.findings-eacl.80

Mijatović, A., Ursić, L., Buljan, I., & Marušić, A. (2023). A pretrained language model for classification of Cochrane Plain Language Summaries on conclusiveness of recommendations [Conference poster]. Faculty of Humanities and Social Sciences, University of Split, Londres.

Ong, E., Jerwin Damay, Lojico, G., Lu, K., & Dex Tarantan. (2008). Simplifying Text in Medical Literature. *Journal of Research in Science Computing and Engineering*, *4*(1). https://doi.org/10.3860/jrsce.v4i1.441

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Finn, C., & Zhang, T. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

Rahman, M. M., Irbaz, M. S., North, K., Williams, M. S., Zampieri, M., & Lybarger, K. (2024). Health Text Simplification: An Annotated Corpus for Digestive Cancer Education and Novel Strategies for Reinforcement Learning. *ArXiv.org.* https://arxiv.org/abs/2401.15043

Ratzan, S., Parker, R., Selden, C., & Zorn, M. (2000). National Library of Medicine Current Bibliographies in Medicine: Health Literacy. Bethesda, MD: National Institutes of Health.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv.org.* https://arxiv.org/abs/1910.01108

Sørensen, K., Pelikan, J. M., Röthlin, F., Ganahl, K., Slonska, Z., Doyle, G., Fullam, J., Kondilis, B., Agrafiotis, D., Uiters, E., Falcon, M., Mensing, M., Tchamov, K., Broucke, S. van den, & Brand, H. (2015). Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *The European Journal of Public Health*, 25(6), 1053–1058. https://doi.org/10.1093/eurpub/ckv043

Zha, Y., Yang, Y., Li, R., & Hu, Z. (2023). AlignScore: Evaluating Factual Consistency with a Unified Alignment Function. *ArXiv.org*. https://arxiv.org/abs/2305.16739

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. *International Conference on Machine Learning*, PMLR, 11393–11403.

Zhang, T., Kishore, V., Wu, F., Weinberger, K., & Artzi, Y. *BERTSCORE: EVALUATING TEXT GENERATION WITH BERT. ICLR 2020.* https://arxiv.org/pdf/1904.09675

Zheng, L., Huang, Y., He, J., Zhang, M., & Sun, B. (2023). LLM-as-a-Judge: Evaluating LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Anexos

Anexo 1. Prompt C1 de Arias-Russi et al. (2025).

Using the following abstract of a biomedical study as input, generate a Plain Language Summary (PLS) understandable by any patient, regardless of their health literacy. Ensure that the generated text adheres to the following instructions which should be followed step-by-step:
a. Specific Structure: The generated PLS should be presented in a logical order, using the following order:

1. Plain Title
2. Rationale
3. Trial Design
4. Results

b. Sections should be authored following these parameters:

1. **Plain Title:** Simplified title understandable to a layperson that summarizes the research that was done.
2. **Rationale:** Include: background or study rationale providing a general description of the condition, what it may cause or why it is a burden for the patients; the reason and main hypothesis for the study; and why the study is needed, and why the study medication has the potential to treat the condition.
3. **Trial Design:** Answer 'How is this study designed?' Include the description of the design, description of study and patient population (age, health condition, gender), and the expected amount of time a person will be in the study.
4. **Results:** Answer 'What were the main results of the study', include the benefits for the patients, how the study was relevant for the area of study, and the conclusions from the investigator.

c. Consistency and Replicability: The generated PLS should be consistent regardless of the order of sentences or the specific phrasing used in the input protocol text.
d. Compliance with Plain Language Guidelines: The generated PLS must follow all these plain language guidelines:

- Have readability grade level of 6 or below.
- Do not have jargon. All technical or medical words or terms should be defined or broken down into simple and logical explanations.
- Active voice, not passive.
- Mostly one or two syllable words.
- Sentences of 15 words or less.
- Short paragraphs of 3-5 sentences.
- Simple numbers (e.g., ratios, no percentages).

e. Do not invent Content: The AI model should not invent information. If the AI model includes data other than the one given in the input abstract, the AI model should guarantee such data is verified and real.
f. Aim for an approximate PLS length of 500-900 words.

Anexo 2. MetaScore.

El MetaScore es una métrica compuesta diseñada para evaluar la calidad global de textos generados por modelos LLM, integrando múltiples dimensiones de calidad lingüística y factualidad. En lugar de depender de una única medida (p. ej., BLEU, ROUGE o perplexity), el MetaScore combina seis indicadores complementarios:

1. **Relevancia (BERTScore):** Evalúa la consistencia entre el texto generado y la respuesta esperada utilizando embeddings contextualizados.
2. **Alineamiento semántico (AlignScore-Factuality):** Mide el grado de correspondencia conceptual entre la generación y la fuente original o resumen de referencia, típicamente empleando modelos embedding especializados.
3. **Legibilidad (Readability Index):** Basado en métricas de complejidad y estructura lingüística (e.g., Flesch–Kincaid), estima cuán accesible es el texto producido.
4. **Coherencia (Coherence Score):** Evalúa la continuidad semántica y la progresión lógica entre frases del texto generado.
5. **Concisión (Conciseness Score):** Mide en qué grado el texto es informativo y libre de redundancias respecto a la información fuente.
6. **Perplejidad (PPL):** Indica la fluidez probabilística y la adecuación lingüística de acuerdo con un modelo de lenguaje independiente.

El MetaScore combina estas dimensiones mediante una función agregadora que normaliza y pondera cada sub-métrica, permitiendo una evaluación holística útil para comparar modelos, analizar configuraciones de entrenamiento (p. ej., SFT, RLHF, LoRA) y diagnosticar degradaciones en la calidad generativa. Esta aproximación ha sido adoptada recientemente en evaluaciones internas de varios LLM para capturar simultáneamente fidelidad semántica, calidad estilística y robustez lingüística.