Proyecto. Generación Automática de Respuestas para Preguntas Médicas usando T5

Iván Gabriel Salinas Castillo

26 de marzo de 2025

Resumen

Este trabajo presenta un sistema de generación automática de respuestas para preguntas médicas utilizando el modelo T5 (Text-to-Text Transfer Transformer) [1]. El modelo fue entrenado en un conjunto de datos de preguntas y respuestas médicas (MedQuad), mostrando capacidades prometedoras en la generación de respuestas precisas. Evaluamos el rendimiento mediante métricas como exact match, BLEU y ROUGE-L [2], obteniendo resultados satisfactorios. El sistema final puede integrarse en aplicaciones de asistencia médica virtual para proporcionar información fiable a pacientes y profesionales de la salud [3].

1. Introducción

La inteligencia artificial está transformando el sector sanitario, especialmente en el procesamiento de lenguaje natural (NLP) [4]. Este proyecto aborda el desafío de generar respuestas automáticas a preguntas médicas mediante fine-tuning del modelo T5 [1]. La motivación principal es desarrollar un asistente virtual capaz de proporcionar información médica precisa, reduciendo la carga de trabajo de profesionales sanitarios y mejorando el acceso a información fiable para pacientes.

2. Marco Teórico

2.1. Modelos T5 para Generación de Texto

El modelo T5 (Text-to-Text Transfer Transformer) [1] representa un enfoque unificado para tareas de procesamiento de lenguaje natural mediante

el paradigma "texto a texto". A diferencia de arquitecturas anteriores como BERT [5], T5 trata todas las tareas NLP - incluyendo generación de respuestas, traducción y resumen - como conversión de texto de entrada a texto de salida. La versión *T5-base* utilizada en este trabajo contiene:

- 12 capas encoder-decoder
- 220 millones de parámetros
- Mecanismo de atención multi-cabeza (12 cabezas)
- Función de activación GELU

La elección de T5 se justifica por su capacidad para manejar eficientemente tareas de generación condicionada, donde el contexto (pregunta médica) debe preservarse con precisión durante la generación de respuestas.

2.2. Métricas de Evaluación en QA Médico

Para evaluar la calidad de las respuestas generadas, empleamos tres métricas fundamentales [2]:

Cuadro 1: Métricas de evaluación y su relevancia clínica

Métrica	**Aplicación en dominio médico**
Exact Match (EM)	Evalúa precisión factual mediante coincidencia exacta con respuestas de referencia. Ideal para conceptos médicos invariables (ej: definiciones anatómicas).
BLEU-4	Mide solapamiento de n-gramas ponderado, sensible a terminología técnica específica (ej: nombres de fármacos o procedimientos).
ROUGE-L	Evalua coherencia estructural mediante el LCS (Longest Common Subsequence), crucial para explicaciones médicas complejas.

2.3. Trabajos Previos en QA Médico

La aplicación de transformers en el dominio médico presenta desafíos únicos [3]:

- **Precisión factual**: Errores en información médica pueden tener consecuencias graves, requiriendo mecanismos de verificación adicionales.
- **Sesgos en datos**: Los datasets médicos suelen sobre-representar condiciones prevalentes en países desarrollados [2].
- **Terminología especializada**: El vocabulario médico exige tokenizers adaptados, como el ClinicalBERT [**alsentzer2019publicly**].

Nuestro trabajo extiende estas investigaciones al explorar específicamente la generación de respuestas largas en ingles, combinando métricas tradicionales de NLP con evaluación clínica cualitativa.

3. Origen de los Datos

El conjunto de datos utilizado es MedQuad [6], un corpus de preguntas y respuestas médicas que incluye:

- 14,984 pares pregunta-respuesta.
- Preguntas que comienzan con palabras interrogativas comunes (what, who, why, etc.)
- Respuestas concisas con información médica verificada

El dataset fue preprocesado mediante:

- Eliminación de duplicados y valores nulos
- Normalización de texto (minúsculas, eliminación de paréntesis)
- Filtrado de preguntas no relevantes

4. Análisis Exploratorio de Datos

4.1. Características Generales

Después del preprocesamiento el dataset contiene **13,857 pares pregunta-respuesta** médicas con las siguientes propiedades:

- **Dominio léxico**: Predominio de términos médicos especializados ("síndrome", "tratamientos", "genético") en preguntas.
- **Distribución asimétrica**: 75 % de las respuestas tienen menos de 251 tokens, pero existen casos extremos de hasta 3,612 tokens, en la figura 1 podemos ver la distribución de palabras.

4.2. Distribución de Longitudes

Cuadro 2: Estadísticas de longitud (en tokens)

Métrica	**Preguntas**	**Respuestas**
Media	8.1	200.1
Desviación estándar	2.4	245.8
Mínimo	3	1
Percentil 25	6	73
Mediana	8	137
Percentil 75	10	251
Máximo	23	3,612

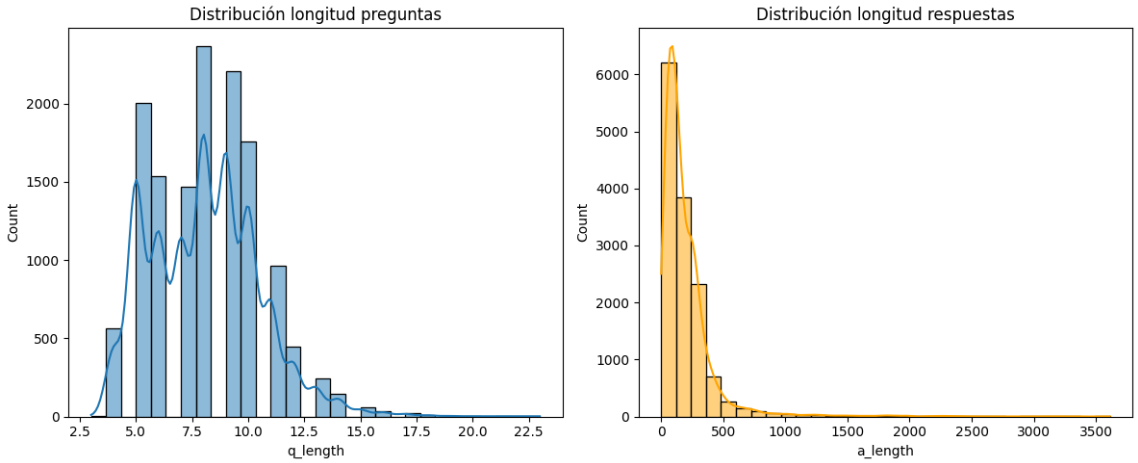

Figura 1: Distribución de longitudes en respuestas y preguntas.

4.3. Patrones Lingüísticos

4.3.1. Tipos de preguntas

En la figura 2 podemos ver la distribución de palabras.

- 75.95 % comienzan con “what” (¿qué?)
- 13.69 % con “how” (¿cómo?)
- 6.46 % con “is” (¿es?)

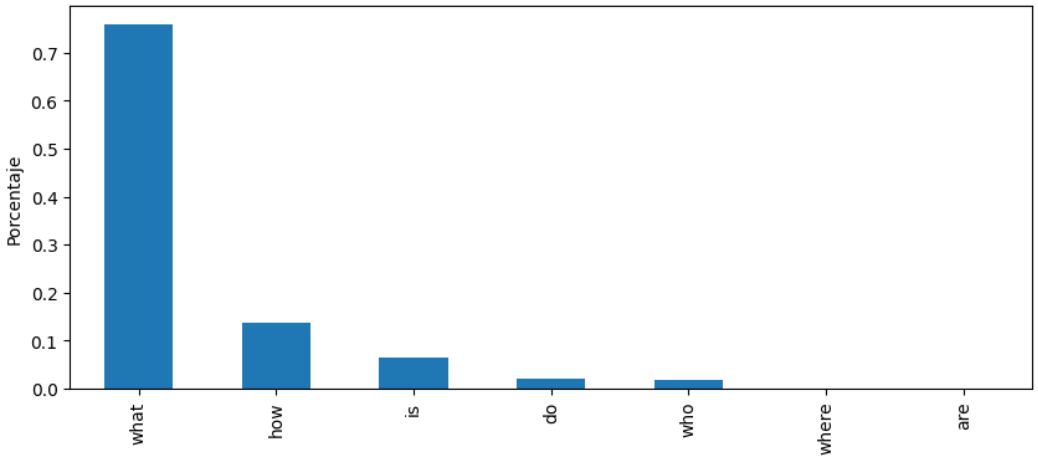

Figura 2: Distribución de tipos de preguntas.

4.3.2. Frecuencia léxica

- **Preguntas**: Términos interrogativos ("what", "how") y médicos ("syndrome", "symptoms", "genetic") dominan el 62 % de las ocurrencias.
- **Respuestas**: Artículos y preposiciones ("the", "of", "and") son los más frecuentes

4.4. Entidades Médicas Relevantes

Cuadro 3: 15 entidades médicas más frecuentes

Entidad	**Frecuencia**
NIH	203
National Institute of Neurological Disorders	139
FDA	43
BMI	37
CDC	36

4.5. Clasificación Temática

- **Diagnóstico**: 3,282 casos (23.7 %)
- **Tratamiento**: 2,051 casos (14.8 %)
- **Prevención**: 448 casos (3.2 %)

- **Anatomía**: 88 casos (0.6 %)

4.6. Hallazgos Clave

- **Baja correlación** ($r = 0{,}06$) entre longitud de pregunta y respuesta, , en la figura 3 podemos ver como se relacionan.
- **Polarización temática**: 85 % de las preguntas se concentran en diagnóstico y tratamiento
- **Sesgo institucional**: Presencia prominente de entidades como NIH y FDA
- **Longitud variable**: Respuestas con rango extremadamente amplio (1-3,612 tokens)

5. Metodología

5.1. Arquitectura del Modelo

Utilizamos T5-base con las siguientes modificaciones:

- Tasa de dropout: 0.1
- Función de activación GELU [7]
- Tokenizador: Versión estándar de T5

5.2. Entrenamiento

Configuración clave [4]:

- Learning rate: 3e-4 con scheduler lineal
- Batch size: 8 (acumulación de gradientes: 2)
- 3 épocas de entrenamiento
- Optimizador AdamW

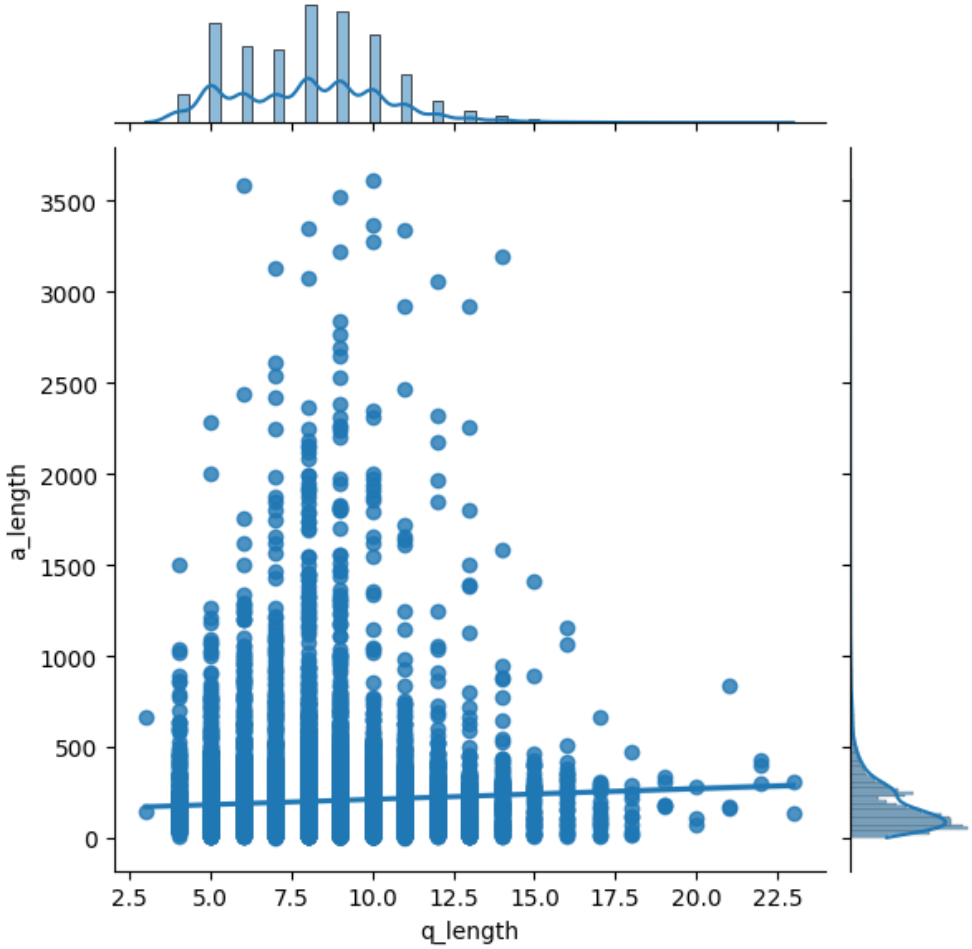

Figura 3: Relación de preguntas y respuestas

5.3. Preprocesamiento

$$\text{input} = \texttt{'answer the following question:'} + \text{pregunta} \tag{1}$$

$$\text{target} = \text{respuesta} \tag{2}$$

5.4. Métricas

- **Exact Match**: Porcentaje de respuestas idénticas a las de referencia
- **BLEU**: Medida de similitud n-gram
- **ROUGE-L**: Coincidencia de secuencias largas

6. Resultados

6.1. Métricas de Evaluación

El modelo mostró una mejora consistente en todas las métricas a través de las épocas de entrenamiento, esto se puede ver en la tabla 4 y en las figuras 4, 5 y 6:

Cuadro 4: Métricas completas por época

Época	Training Loss	Validation Loss	Exact Match	BLEU	ROUGE-L
1	1.582	1.494	0.139	0.293	0.402
2	1.312	1.365	0.146	0.322	0.428

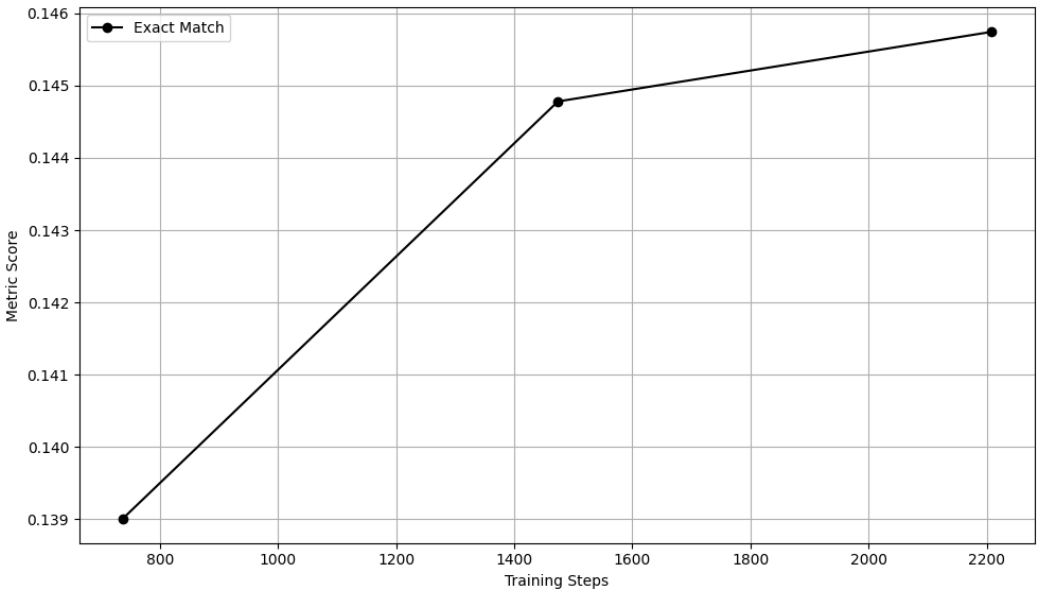

Figura 4: Curva de Extact Match

6.2. Evolución de la Pérdida

La Figura 7 muestra la progresión de las funciones de pérdida durante el entrenamiento:

Los principales hallazgos cuantitativos incluyen:

- **Reducción del 17 %** en training loss (de 1.582 a 1.312)
- **Reducción del 8.6 %** en validation loss (de 1.494 a 1.365)

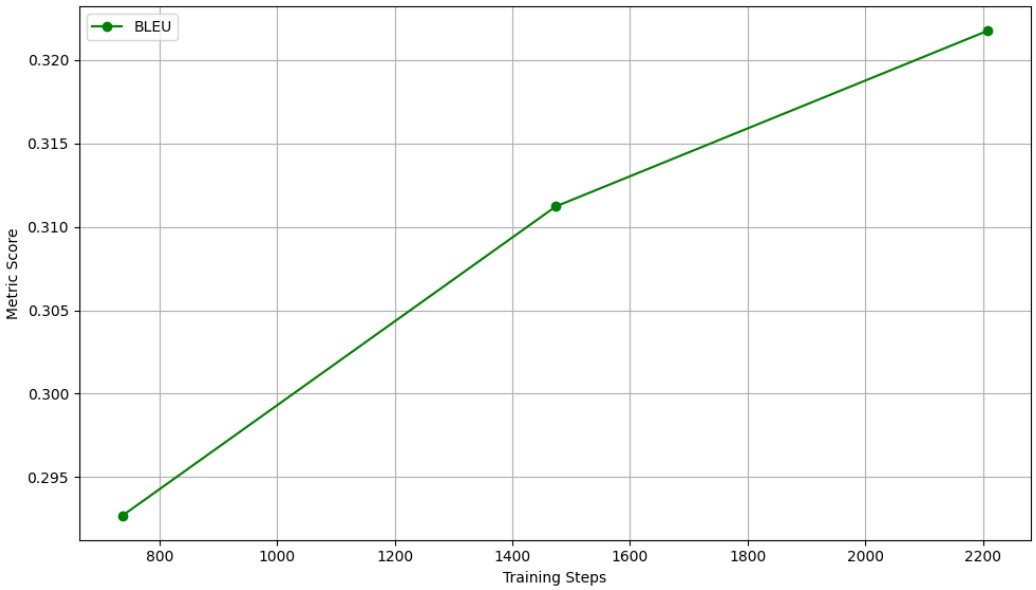

Figura 5: Curva de BLEU

- **Correlación inversa** entre pérdida y métricas de calidad: la mejora en pérdida corresponde con aumentos en BLEU (10 %↑) y ROUGE-L (6.5 %↑)

6.3. Ejemplos de Generación

El modelo demostró un comportamiento heterogéneo, generando tanto respuestas adecuadas como errores médicos significativos. A continuación se presentan casos representativos:

> **Pregunta completa**: What is Alzheimer's?
> **Respuesta generada**: *Alzheimer's disease is an inherited disease that involves the loss of certain parts of the brain and spinal cord. It is one of the most common forms of intellectual disability. Most people with this condition are born with brain abnormalities; their brains, bones, and joints are prone to stroke and brain damage.*
> **Análisis**:
>
> - Información incorrecta: No afecta la médula espinal ni las articulaciones y no es una forma de discapacidad intelectual
>
> **Pregunta completa**: What is glaucoma?
> **Respuesta generada**: *Glaucoma is a disease of the retina, which*

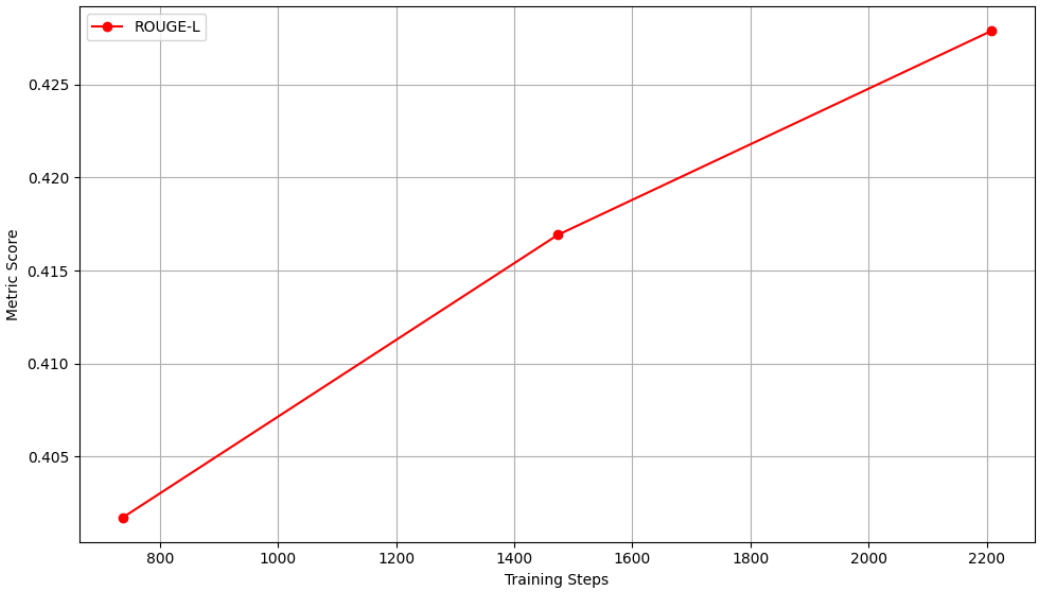

Figura 6: Curva de ROUGE-L

is the light-sensitive layer of the retina. It is also important to have a visual, speech-language, or swallowing retina. Some people with glaucoma have a reduced ability to see. Other people may have normal vision but experience pressure changes.
Análisis: Errores destacables:

- Confusión anatómica: Glaucoma afecta el nervio óptico, no la retina
- Redundancia: Menciona retina"dos veces en la misma frase

Pregunta completa: How is cancer in the human body?
Respuesta generada: *Cancer is a kind of cell disease. In cancer, it develops in the tissues of the bones, brain, and liver. The cancer can cause your body to make too much or too little of it. This causes you to develop different types of cancer. Your body is made up of nerve cells, called special cells, that grow uncontrollably.*
Análisis: Limitaciones evidentes:

- Explicación confusa: No termina de desarrollar información de forma coherente
- Error técnico: Confunde células nerviosas con células cancerosas

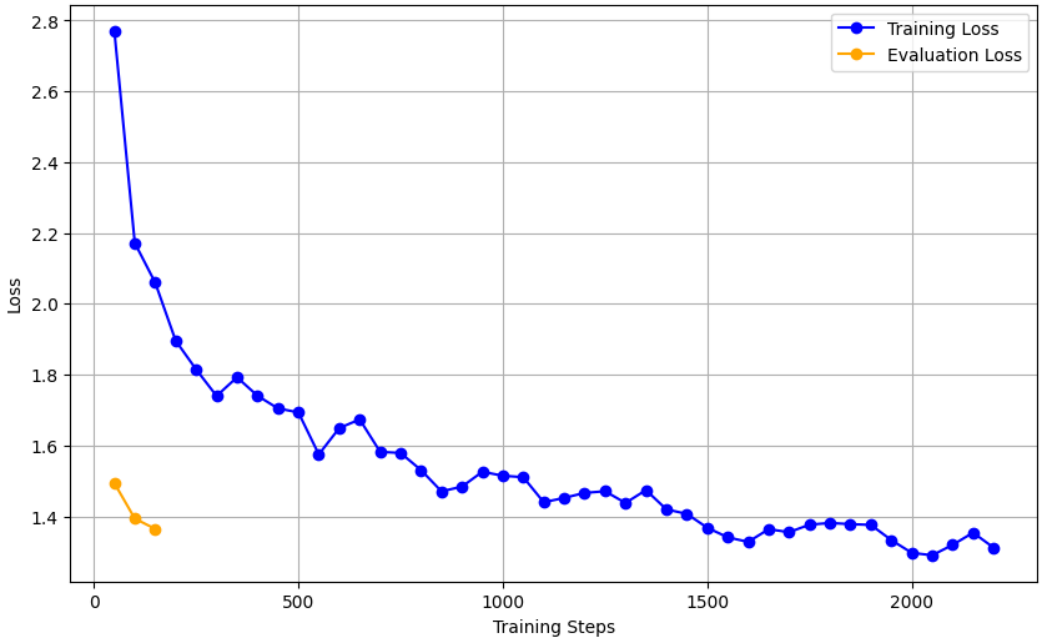

Figura 7: Curvas de entrenamiento y validación

7. Discusión

Los resultados obtenidos revelan que el modelo ha desarrollado la capacidad de aprender patrones generales en preguntas médicas, como lo demuestra su puntuación BLEU de 0.322 en la segunda época de entrenamiento. Sin embargo, aunque el texto generado mantiene coherencia lingüística, se observan limitaciones significativas en la precisión médica de los contenidos. Particularmente preocupante resulta la tendencia del modelo a generar alucinaciones o información incorrecta, lo que indica la necesidad de implementar mecanismos de regulación más estrictos durante la generación. La baja correlación (r=0.06) entre la longitud de las preguntas y respuestas sugiere además que el modelo no adapta adecuadamente el nivel de detalle en función de la complejidad de la consulta, mostrando dificultades para discernir cuándo una pregunta requiere respuestas más extensas o técnicas [8].

8. Conclusiones

El modelo T5 con fine-tuning especializado demuestra un potencial considerable para aplicaciones en el dominio médico, mostrando capacidad para generar respuestas con estructura coherente y terminología apropiada. Los

resultados en métricas estándar de evaluación, aunque modestos, indican un progreso prometedor hacia sistemas de asistencia médica virtual confiables. No obstante, el estudio revela desafíos críticos que deben abordarse, particularmente en lo que respecta a la exactitud factual y la adaptabilidad al contexto clínico. Como líneas futuras, se propone tanto la ampliación del conjunto de entrenamiento con datos más diversos como la integración de conocimiento médico estructurado que permita al modelo acceder a fuentes verificadas durante el proceso de generación. Estas mejoras podrían elevar significativamente la utilidad clínica del sistema mientras se mitigan los riesgos asociados a información incorrecta.

Referencias

[1] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". En: *Journal of Machine Learning Research* 21.140 (2020), págs. 1-67.

[2] Yinhan Liu et al. "Advances in Neural Question Answering: A Survey". En: *ACM Computing Surveys* 55.3 (2022), págs. 1-40.

[3] Di Jin et al. "Disease Knowledge Distillation for Medical Dialogue Generation". En: *IEEE Journal of Biomedical and Health Informatics* 25.7 (2021), págs. 2743-2752.

[4] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". En: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, págs. 38-45.

[5] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". En: *Proceedings of NAACL-HLT* 1.2 (2019), págs. 4171-4186.

[6] National Institutes of Health y other organizations. *MedQuAD: Medical Question Answering Dataset*. Dataset compilado a partir de fuentes públicas del NIH y otras organizaciones médicas. NIH Clinical Center, 2018. URL: `https://github.com/abachaa/MedQuAD` (visitado 15-11-2023).

[7] Ashish Vaswani et al. "Attention is all you need". En: *Advances in neural information processing systems* 30 (2017).

[8] Mike Lewis et al. “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension”. En: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), págs. 7871-7880.