UNIVERSIDAD EAFIT

Ajuste fino de un modelo LLM para realizar reportes resumidos de expertos en trading, con integración de datos desde redes sociales

Fine-tuning an LLM Model to Generate Summarized Expert Trading Reports, with Integration of Data from Social Networks

Andres Felipe Restrepo Acevedo

Proyecto de grado:

Asesor, docente

Juan David Martínez Vargas

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA

MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA

MEDELLÍN

2025

Resumen

El mercado financiero contemporáneo se caracteriza por su alta complejidad y el volumen masivo de datos estructurados y no estructurados que genera diariamente, lo que representa un desafío significativo para los inversores individuales en cuanto al análisis y la toma de decisiones informadas. Este proyecto propone el ajuste fino de un modelo de lenguaje pequeño (SLM, por sus siglas en inglés) integrado en una herramienta capaz de generar reportes de análisis financiero similares a los elaborados por expertos. Para la prueba de concepto (PoC), se emplean transcripciones de videos de análisis financiero publicados por expertos en sus canales de YouTube. El modelo SLM es ajustado mediante técnicas de fine-tuning con instrucciones específicas y la incorporación de la técnica LoRa (Low-Rank Adapters), con el objetivo de extraer y resumir información clave relevante para los inversores individuales. El propósito principal de esta herramienta es asistir a los inversores individuales mediante la generación de reportes eficientes y accesibles, facilitando el acceso a información valiosa en lenguaje natural y mejorando su capacidad para tomar decisiones fundamentadas a partir de datos no estructurados, todo ello con una inversión mínima de tiempo y recursos.

Los resultados experimentales demuestran la viabilidad de utilizar Modelos de Lenguaje Pequeños (SLMs) ajustados para la generación automatizada de reportes financieros de calidad. Específicamente, el modelo finetune_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v2_Q8_0 seleccionado alcanzó una puntuación promedio de 5,67 sobre 10 en la evaluación realizada por un LLM evaluador, con una distancia de coseno promedio de 0,159 respecto a los resúmenes de referencia generados por el modelo preentrenado fundacional GPT-4.1. Esta mejora representa un incremento del 97,5% en el rendimiento en comparación con el modelo base Llama 3.1 8B Instruct sin ajuste fino. Cualitativamente, el modelo exhibe una alta fidelidad y coherencia en la extracción y síntesis de información clave en contextos de longitud moderada, aunque presenta desafíos en la interpretación temática para transcripciones considerablemente extensas. Adicionalmente, la implementación de esta herramienta proyecta un ahorro anual estimado de 560 horas para inversores individuales, junto con una reducción anual de costos de API estimada entre 7,52 y 25 dólares para los canales analizados en la prueba de concepto.

Palabras clave: Modelos de lenguaje grandes (LLM), Destilación de conocimiento, Ajuste fino, LoRa (Low-Rank Adaptation), Evaluación con LLM, YouTube.

Abstract

The contemporary financial market is characterized by its high complexity and the massive volume of structured and unstructured data generated daily, posing significant challenges for individual investors in terms of analysis and informed decision-making. This project proposes the fine-tuning of a Small Language Model (SLM) integrated into a tool capable of generating financial analysis reports similar to those produced by experts. For the proof of concept (PoC), transcripts from financial analysis videos published by experts on their YouTube channels are utilized. The SLM is fine-tuned using instruction-based techniques and the incorporation of the LoRa (Low-Rank Adapters) method, with the aim of extracting and summarizing key information relevant to individual investors. The main objective of this tool is to assist individual investors by generating efficient and accessible reports, facilitating access to valuable information in natural language, and enhancing their ability to make data-driven decisions from unstructured data, all with minimal investment of time and resources.

Experimental results demonstrate the viability of using fine-tuned Small Language Models (SLMs) for the generation of high-quality financial reports. Specifically, the selected model, finetune_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v2_Q8_0, achieved an average score of 5.67 out of 10 in the evaluation conducted by a judge LLM, with an average cosine distance of 0.159 compared to the reference summaries generated by the foundational pretrained model GPT-4.1. This improvement represents a 97.5% increase in performance compared to the same base model, Llama 3.1 8B Instruct, without fine-tuning. Qualitatively, the model exhibits high fidelity and coherence in the extraction and synthesis of key information in moderately long contexts, although it faces challenges in thematic interpretation when dealing with considerably lengthy transcripts. Additionally, implementation of this tool is projected to save 560 hours annually for individual investors, along with an estimated annual reduction in API costs ranging from $7.52 to 25$ for the channels analyzed in the proof of concept.

Keywords: Large Language Models (LLM), Knowledge Distillation, Fine-tuning, LoRa (Low-Rank Adaptation), LLM as Evaluator, YouTube.

Contenido

Contents

Lista de Figuras

List of Figures

Lista de Tablas

List of Tables

1 Introducción

El sector financiero maneja grandes cantidades de datos no estructurados que son difíciles de procesar y analizar de manera eficiente, lo que limita la capacidad de tomar decisiones informadas y de innovar en productos y servicios financieros. El uso de técnicas avanzadas de procesamiento de lenguaje natural (PLN) permite transformar estos datos no estructurados en información útil, facilitando la toma de decisiones basada en datos y fomentando la innovación en el sector financiero (Du et al., 2025). El mercado financiero actual es un entorno dinámico y complejo, caracterizado por un flujo constante de información proveniente de diversas fuentes, como redes sociales, medios de comunicación y plataformas especializadas en análisis financiero. Esta gran cantidad de datos ofrece una oportunidad para los inversores, quienes pueden acceder a información valiosa de manera instantánea. Sin embargo, el desafío radica en procesar y analizar eficazmente este gran volumen de datos para tomar decisiones de inversión informadas. Según Pinto et al. (2021), la capacidad de los inversores individuales para gestionar tanta información es limitada, lo que a menudo resulta en decisiones basadas en información incompleta o incorrecta.

La creciente interdependencia entre el análisis fundamental, técnico y de sentimiento en las redes sociales ha generado la necesidad de herramientas más sofisticadas que puedan integrar datos de diversas fuentes para ofrecer una visión más holística del mercado. Esta integración es crucial, ya que no solo permite analizar los movimientos del mercado en función de los datos históricos, sino también comprender el impacto del sentimiento de los inversores en tiempo real (Syamala Rao M. et al., 2023). Además, los inversores individuales, a diferencia de las grandes instituciones, carecen de recursos avanzados para procesar y contextualizar rápidamente esta información como la desarrollada por la empresa Infront Analytics en su plataforma de servicios de análisis (Grauer, 2024), lo que puede afectar su capacidad de reaccionar ante cambios en el mercado de manera oportuna (Ji et al., 2021).

Dada esta problemática, surge la necesidad de desarrollar herramientas que puedan sintetizar información financiera de diversas fuentes, incluyendo videos de YouTube, redes sociales, datos de mercado y pronósticos de modelos tradicionales o de aprendizaje automático (ML) en tiempo oportuno. Utilizando técnicas avanzadas como el procesamiento de lenguaje natural (NLP) y modelos de lenguaje grande (LLMs), que han demostrado mejorar significativamente la eficiencia y efectividad en aplicaciones financieras (Dong et al., 2024). Estas tecnologías permiten no solo extraer información clave de fuentes no estructuradas, sino también resumir y contextualizar datos de manera eficiente, facilitando una toma de decisiones más rápida y precisa para los inversores individuales (Ji et al., 2021; S. Mukherjee et al., 2023). A pesar de los avances, la síntesis de transcripciones de fuentes de videos de expertos de forma costo eficiente y oportuna para soportar análisis de inversores individuales sigue siendo un desafío no completamente resuelto.

El desarrollo de una herramienta, que integre los componentes mencionados anteriormente, tiene un gran potencial para transformar el modo en que los inversores acceden y utilizan la información financiera. Este tipo de herramientas puede ofre-

cer reportes personalizados en tiempo oportuno, mejorando significativamente la capacidad de los usuarios para tomar decisiones estratégicas en un entorno de alta volatilidad (D. Chen, 2023).

La metodología para este proyecto incluye la recolección de análisis realizados por expertos a partir de transcripciones de videos de YouTube de canales especializados en análisis técnico/fundamental de activos financieros seleccionados. Se realizan resúmenes de las transcripciones de los videos, mediante un modelo SLM ajustado para la tarea a partir de conjuntos de datos generados previamente con LLM fundacional pre-entrenado como gpt4.1, con el objetivo de optimizar costos en el sistema y que este pueda ser implementado en un cómputo local o en la nube consumiendo menores recursos y sin limitaciones de consumo mensuales (límites de API openIA). Se desplegará el modelo sobre una aplicación web que permita generar reportes resumidos de videos de YouTube de análisis de expertos en trading.

Este proyecto tendrá un impacto significativo en la forma en que los inversores individuales acceden a análisis de expertos en tiempo oportuno, aportando información valiosa para soportar la toma de decisiones en los mercados financieros. Además, la utilización de SLMs para la generación de resúmenes y la extracción de información de activos permite optimizar los costos de implementación del sistema. En conjunto, la implementación de esta herramienta no solo optimiza recursos, sino que representa una aplicación tangible de la inteligencia artificial y la ciencia de datos para democratizar y facilitar el acceso a información financiera experta, brindando a los inversores individuales una solución práctica y efectiva para mejorar sus decisiones de inversión.

2 Planteamiento del problema

El entorno financiero contemporáneo se caracteriza por su alta complejidad y un volumen masivo de datos estructurados y no estructurados que se generan diariamente. Este escenario presenta una serie de desafíos significativos para los inversores individuales, quienes a menudo se enfrentan a dificultades para procesar, analizar y tomar decisiones informadas con base en la información financiera disponible. Dentro de este vasto panorama, los videos de análisis publicados por expertos en plataformas como YouTube, aunque representan una fuente particularmente rica por su profundidad y formato, presentan a su vez desafíos específicos para la extracción y síntesis eficiente de su contenido, especialmente para el inversor individual. La problemática central se manifiesta a través de un conjunto de factores interrelacionados que limitan la capacidad del inversor individual para operar eficazmente en este mercado dinámico:

Causas subyacentes del desafío:

- El flujo masivo de información proveniente de diversas fuentes, incluyendo noticias, informes de ganancias, comentarios en redes sociales y análisis de expertos, resulta en una sobrecarga de datos para el inversor individual(Onan and Dursun, 2024).
- La creciente interdependencia entre el análisis fundamental, técnico y de sentimiento en las redes sociales ha generado la necesidad de herramientas más sofisticadas que puedan integrar datos de diversas fuentes para ofrecer una visión holística del mercado (Syamala Rao M. et al., 2023).
- La volatilidad del sentimiento del mercado, reflejada en el contenido generado por usuarios en internet, tiene un impacto significativo en el comportamiento a corto plazo del mercado de valores (Nguyen et al., 2015; Ren et al., 2019).
- La naturaleza inherente del lenguaje hablado en las transcripciones de videos de análisis, que tiende a ser menos estructurado, incluyendo frases incompletas, repeticiones, cambios abruptos de tema, jerga técnica y referencias a contexto visual, complica significativamente su interpretación automática y la identificación de puntos clave (D. Chen, 2023; S. Mukherjee et al., 2023).

Efectos directos en la toma de decisiones del inversor individual:

- La capacidad limitada del inversor individual para procesar y evaluar la vasta cantidad de información disponible aumenta el riesgo de tomar decisiones basadas en datos incompletos o desactualizados (Onan and Dursun, 2024; Pinto et al., 2021).
- La dificultad para cuantificar y analizar el sentimiento del mercado de manera precisa y rápida incrementa el reto de reaccionar oportunamente ante las fluctuaciones bursátiles (Onan and Dursun, 2024).

- La ausencia de herramientas que proporcionen una síntesis eficiente de datos en tiempo real, permitiendo la contextualización rápida de la información con tendencias históricas y análisis fundamentales, limita la capacidad de toma de decisiones informadas (D. Chen, 2023; S. Mukherjee et al., 2023).

Barreras tecnológicas y de acceso a soluciones eficientes:

- Los métodos tradicionales de análisis financiero, basados en estadísticas y herramientas numéricas sencillas, son insuficientes para afrontar la creciente complejidad y el volumen de datos en los mercados financieros actuales (Onan and Dursun, 2024; Pinto et al., 2021).

- Los inversores individuales suelen carecer de los recursos y tecnologías avanzados que poseen las grandes instituciones para realizar un análisis exhaustivo y eficaz, lo que afecta su capacidad de reaccionar ante cambios en el mercado de manera oportuna (Ji et al., 2021; Pelster and Val, 2024).

- La extracción de insights valiosos de fuentes no estructuradas como transcripciones de videos requiere técnicas avanzadas de Procesamiento de Lenguaje Natural (PLN), las cuales no están fácilmente disponibles o son costo-eficientes para el inversor individual (S. Mukherjee et al., 2023).

- El alto costo computacional y de implementación de los Modelos de Lenguaje Grandes (LLMs) de última generación, con cientos de millones de parámetros, limita su uso generalizado en aplicaciones reales para el inversor individual (Turc et al., 2019; Xu et al., 2023).

En resumen, el problema central radica en la ausencia de herramientas accesibles, costo-eficientes y efectivas que permitan a los inversores individuales acceder, resumir y contextualizar, en tiempo oportuno, la información clave contenida en análisis financieros generados por expertos, particularmente en formatos no estructurados como videos. Estos desafíos motivan el diseño de una herramienta basada en modelos de lenguaje pequeño (SLM) ajustados, capaz de ofrecer reportes personalizados y resumidos que ayuden al inversor individual a tomar decisiones informadas de forma eficiente, democratizando así el acceso a información financiera experta y optimizando el tiempo y los recursos.

3 Justificación

El presente trabajo aborda una problemática crítica en el mercado financiero contemporáneo: la dificultad inherente de los inversores individuales para procesar y analizar la vasta cantidad de información que se genera diariamente de manera oportuna. La sobrecarga de datos no estructurados, provenientes de múltiples fuentes como redes sociales, videos de análisis de expertos y noticias, se erige como un obstáculo significativo que limita la capacidad de los inversores para tomar decisiones informadas y estratégicas. Los métodos tradicionales de análisis financiero resultan insuficientes ante la creciente complejidad y el volumen de información actual. Este proyecto propone el ajuste fino de un Modelo de Lenguaje Pequeño (SLM) para generar reportes resumidos de análisis de expertos, con el fin de proporcionar insights actualizados.

La relevancia de esta investigación radica en su capacidad para democratizar el acceso a análisis financieros especializados, contribuyendo a reducir la asimetría de información y la brecha existente entre los inversores institucionales, que disponen de herramientas avanzadas, y los inversores individuales, quienes carecen de recursos similares para interpretar grandes volúmenes de datos. Al ofrecer una solución basada en inteligencia artificial que sintetiza información clave, este sistema no solo mejora la capacidad de los inversores individuales para tomar decisiones fundamentadas en un entorno volátil y cambiante, sino que también les proporciona un ahorro significativo de tiempo, estimado en un promedio de 560 horas anuales para el análisis de los canales seleccionados en este estudio. El valor agregado y el enfoque innovador de esta solución se fundamentan en la combinación estratégica de técnicas avanzadas de Procesamiento de Lenguaje Natural (PLN):

- La aplicación del ajuste fino de modelos de lenguaje con bajo consumo de recursos (LoRA), una técnica eficiente que permite adaptar modelos de gran escala como Llama 3.1 sin incurrir en los elevados costos computacionales del full fine-tuning, optimizando así los requisitos de hardware y los costos de implementación del sistema.

- La generación automática de conjuntos de datos sintéticos para el entrenamiento y prueba mediante un modelo fundacional pre-entrenado (GPT-4.1), lo cual ha demostrado mejorar el rendimiento de modelos más pequeños y ofrece un método rentable para su ajuste fino.

- La evaluación de los reportes generados asistida por Modelos de Lenguaje Grandes (LLMs como evaluadores), una metodología que ha demostrado una fuerte correlación con las evaluaciones humanas, superando en muchos casos a las métricas automáticas tradicionales en la valoración de la calidad de textos generados.

Estas características inherentes al diseño metodológico confieren a la herramienta una flexibilidad, escalabilidad y sostenibilidad temporal significativas. Su concepción

permite la integración en plataformas de inversión o educación financiera, proporcionando a los inversores individuales una solución práctica y efectiva para mejorar sus decisiones estratégicas en un entorno altamente competitivo y de alta volatilidad. La optimización de costos mediante el uso de SLMs refuerza su viabilidad para ser implementada en un cómputo local o en la nube con menores recursos, haciéndola una solución práctica y efectiva para el inversor individual.

4 Objetivos

General

Desarrollar una herramienta que proporcione reportes resumidos personalizados a los inversores individuales, utilizando información proporcionada por expertos en trading y análisis financieros en redes sociales como YouTube, a través de videos, para ofrecer apoyo en la toma de decisiones de inversión de manera oportuna.

Específicos

- Integrar y procesar datos de fuentes como YouTube, para proporcionar información relevante y actualizada a los inversores.
- Hacer ajuste fino de modelo SLM para generar reportes resumidos y con activos mencionados de fuentes de datos no estructurados, para apoyar decisiones de inversión.
- Evaluar la generación de los reportes de los modelos ajustados para seleccionar el mejor y usando métricas como distancia de coseno, métricas que utilizan otro modelo generativo para calificar el reporte resumido generado.
- Desplegar el modelo para que sea usado fácilmente por el usuario final.

5 Marco conceptual

5.1 Contexto del Dominio Financiero

5.1.1 El mercado de valores

El mercado de valores constituye una de las principales instituciones dentro de la economía moderna, ya que permite canalizar recursos hacia actividades productivas a través de la compraventa de instrumentos financieros. En este entorno, el valor de los activos financieros se define no sólo por información fundamental y técnica, sino también por datos provenientes de medios tradicionales y digitales (Huang et al., 2022).

5.1.2 Influencia de las redes sociales y medios de comunicación en el mercado

La incorporación de datos de redes sociales y medios de comunicación al análisis del mercado financiero ofrece una nueva oportunidad para mejorar las predicciones de precios de acciones, especialmente mediante el uso de técnicas más avanzadas de Procesamiento de Lenguaje Natural (PLN) y aprendizaje profundo (Pinto et al., 2021). Estos datos alternativos permiten captar percepciones y reacciones colectivas en tiempo real, contribuyendo a una comprensión más profunda de las fluctuaciones del mercado financiero. La relevancia de estos enfoques ha ido en aumento con los avances en modelos de PLN y aprendizaje profundo, que posibilitan el procesamiento eficiente de grandes volúmenes de información no estructurada.

5.1.3 YouTube como fuente de información financiera

La integración de análisis de contenidos provenientes de plataformas como YouTube, en las que se comparten opiniones, análisis y reportes en formato de video, representa una tendencia emergente por su potencial para captar percepciones y reacciones del público en tiempo real, ampliando así el marco de datos relevantes para la comprensión del mercado financiero (Pinto et al., 2021). En la era digital, los videos se han convertido en el nuevo "idioma común" de Internet, con plataformas sociales como YouTube, Facebook, Instagram y X liderando este cambio. La masiva disponibilidad de videos conlleva desafíos significativos en la gestión y extracción de información relevante, así como en la eliminación de redundancias (Paul et al., 2024). Por ello, la elección de YouTube como fuente de datos en el presente estudio responde a la necesidad de desarrollar herramientas que permitan extraer información relevante y concisa de grandes volúmenes de contenido audiovisual, abordando el desafío identificado en el planteamiento del problema y alineando la metodología con las necesidades actuales del análisis financiero.

5.2 Resumen de texto y su aplicación al problema

5.2.1 Introducción al resumen de textos automático

El resumen automático de textos es un área central en el Procesamiento de Lenguaje Natural, cuyo objetivo es condensar información relevante de grandes cantidades de texto en reportes breves, precisos e informativos. Esta tarea se ha abordado tradicionalmente mediante dos enfoques: el extractivo, que selecciona fragmentos textuales del documento fuente, y el abstractivo, que genera contenido original basado en la comprensión semántica del texto (Shakil et al., 2024).

5.2.2 Técnicas de resumen de texto

Dentro de las técnicas extractivas predominan enfoques como el uso de indicadores estadísticos, bag-of-words y algoritmos de grafos para identificar las partes más relevantes de un texto. Sin embargo, estas técnicas presentan limitaciones al no captar adecuadamente la semántica ni el contexto global, especialmente relevantes en dominios complejos como el lenguaje financiero. Por otra parte, el resumen abstractivo se apoya en modelos modernos de lenguaje y arquitecturas avanzadas que permiten la generación de descripciones compactas y contextualmente ricas (Shakil et al., 2024).

5.2.3 Estado del arte en generación abstractiva de resúmenes

Los avances recientes en modelos basados en arquitecturas Transformer han detonado mejoras sustantivas en la generación abstractiva, logrando una mayor capacidad para capturar relaciones complejas dentro del texto, aunque persisten desafíos como mantener la fidelidad factual y la coherencia narrativa, sobre todo en contenidos financieros que requieren alta precisión y actualización constante (Ji et al., 2021).

5.2.4 Evaluación y métricas en resumen de textos

En la evaluación de resúmenes de texto, si bien métricas como ROUGE-N (ROUGE-1 y ROUGE-2) y ROUGE-L son ampliamente utilizadas, especialmente en el resumen abstractivo, estas a menudo no logran capturar las sutilezas de los resúmenes generados, donde las variantes morfológicas y los sinónimos pueden expresar el mismo concepto sin una superposición léxica directa (Saleh et al., 2024). Para abordar esta limitación, la similitud de coseno (Cosine Similarity) emerge como una métrica fundamental para medir la cercanía semántica entre dos piezas de texto, como los resúmenes generados y las transcripciones originales (Han et al., 2012; Sun and Wang, 2024). Esta medida determina el grado en que dos vectores apuntan en la misma dirección, calculando el coseno del ángulo entre ellos(Han et al., 2012; Ijebu et al., 2025). Es ampliamente empleada en el análisis de texto para cuantificar la similitud entre documentos basándose en la frecuencia de palabras o frases, o más profundamente, en las representaciones vectoriales densas o embeddings generadas

por modelos de lenguaje pre-entrenados(Sun and Wang, 2024). Al aprovechar la capacidad de los modelos transformadores para codificar oraciones en vectores de alta dimensión que reflejan su contenido contextual y semántico, la similitud de coseno permite comparaciones precisas de la similitud semántica entre textos, lo cual es crucial para tareas que van más allá del mero solapamiento de palabras, como la detección de paráfrasis y la recuperación de información (Sun and Wang, 2024). En este proyecto, la similitud de coseno, junto con la evaluación mediante Modelos de Lenguaje Grandes (LLM), será utilizada para valorar automáticamente la calidad de los reportes abstractivos generados, buscando una complementación eficiente y escalable a la revisión manual, y asegurando que los resúmenes capturen el significado principal de la información financiera de manera precisa y útil para los inversores individuales.

5.2.5 LLMs como evaluadores

Recientemente, el uso de modelos de lenguaje grandes (LLMs, por sus siglas en inglés) se ha popularizado como método de evaluación automática para textos generados, incluyendo resúmenes. Diversos trabajos han demostrado que las puntuaciones asignadas por LLMs correlacionan fuertemente con las evaluaciones humanas, superando las métricas automáticas tradicionales (Y.-P. Chen et al., 2024), (Panickssery et al., 2024). Por ejemplo, se ha mostrado que los LLMs pueden evaluar la calidad de textos explicativos en sistemas de recomendación de forma precisa y reproducible, considerando atributos como persuasión, transparencia, precisión y satisfacción (Zhang et al., 2024). De esta manera, en el presente trabajo, los LLMs han sido integrados como evaluadores automáticos para valorar la calidad de los reportes abstractivos generados, aprovechando su capacidad de seguimiento de instrucciones, razonamiento contextual y escalabilidad.

5.3 Modelos de lenguaje y técnicas específicas

5.3.1 Modelos de lenguaje grandes (LLMs)

Los modelos de lenguaje de gran escala (LLMs, por sus siglas en inglés) han revolucionado el Procesamiento de Lenguaje Natural al incorporar arquitecturas avanzadas como los modelos de Secuencia a Secuencia (Seq2Seq), mecanismos de atención, mecanismos de copia y redes de punteros. Adicionalmente, los modelos pre-entrenados como BERT, GPT, T5 y BART han alcanzado resultados de estado del arte en la tarea de generación abstractiva de resúmenes, gracias a su entrenamiento con grandes volúmenes de datos textuales y a sus representaciones contextualizadas de palabras y frases (Shakil et al., 2024). Su capacidad de transformar información compleja en reportes concisos resulta esencial para aplicaciones en dominios especializados, como el financiero.

5.3.2 Ajuste fino con instrucciones y Low-Rank Adaptation (LoRA)

El ajuste fino eficiente de LLMs se ha visto favorecido por técnicas como LoRA (Low-Rank Adaptation), que permite adaptar modelos de lenguaje congelando los pesos originales e introduciendo matrices de bajo rango en cada capa. Esta estrategia reduce significativamente el número de parámetros entrenables y los requisitos de hardware, sin incrementar la latencia durante la inferencia. LoRA también es compatible con pesos cuantificados, como en QLoRA, y facilita las transiciones entre tareas especializadas, constituyendo una alternativa eficiente en recursos para el ajuste fino de estos modelos (Gao et al., 2024; Hu et al., 2021).

5.3.3 Destilación de conocimiento y aumento de datos

La destilación de conocimiento es una técnica que permite transferir el conocimiento de un modelo grande (maestro) a uno más pequeño (estudiante), manteniendo la efectividad del modelo en tareas específicas pero con un menor costo computacional. Este método posibilita que los modelos compactos emulen el comportamiento de los modelos avanzados y generen "pseudo-etiquetas" a partir de datos no etiquetados, maximizando así el aprovechamiento de grandes conjuntos de datos no supervisados (Turc et al., 2019; Udagawa et al., 2023). Por otro lado, la aumentación de datos sintéticos contribuye a mejorar la generalización de los modelos mediante la creación de conjuntos de entrenamiento más diversos y robustos. En particular, los modelos avanzados de lenguaje son empleados tanto para la aumentación de datos en comprensión del lenguaje natural (NLU) como en tareas de clasificación y razonamiento, incrementando el rendimiento en escenarios Few-Shot y Zero-Shot. Marcos de trabajo como ZeroGen, SuperGen y S3 eliminan la necesidad de supervisión humana para la generación de datos de entrenamiento en tareas de NLU y NLI (Pieper et al., 2024).

5.3.4 Modelos de lenguaje y destilación (relación entre LLMs y SLMs)

La destilación de conocimiento (Knowledge Distillation, KD) adquiere un rol fundamental en la era de los LLMs, al facilitar la transferencia de capacidades avanzadas de modelos propietarios, como GPT-4, hacia modelos de código abierto como LLaMA y Mistral. Esto permite desarrollar Small Language Models (SLMs) eficientes que heredan competencias avanzadas y pueden ser utilizados en aplicaciones específicas donde los recursos son limitados y la eficiencia es prioritaria (Gu et al., 2024).

5.4 LLMs en el contexto financiero

En los últimos años, los modelos de lenguaje grandes (LLMs) han demostrado ser herramientas disruptivas para el análisis y procesamiento de información financiera. Estos modelos permiten abordar tareas complejas como la predicción de precios de acciones, el análisis automático de informes financieros y la generación de resúmenes de información clave (Lee et al., 2025). Sin embargo, los modelos generales enfrentan

importantes retos en dominios especializados como las finanzas, lo que ha impulsado el desarrollo de Modelos de Lenguaje específicos para Finanzas (FinLLMs) (Lee et al., 2025).

Diversas investigaciones han evaluado la aplicabilidad de los LLMs en el sector financiero. Kirtac and Germano (2024) analizan modelos como BERT, OPT y FinBERT para el análisis de sentimiento en noticias, con resultados que muestran que OPT, basado en GPT-3, supera significativamente otras estrategias en predicción de retornos bursátiles, destacando el potencial transformador de estas tecnologías en la gestión de portafolios. Por su parte, Pelster and Val (2024) examinan cómo ChatGPT puede sintetizar noticias financieras y responder a eventos clave, encontrando una correlación positiva entre las calificaciones generadas por el modelo y los anuncios futuros de resultados y rendimientos bursátiles, lo que sugiere que la IA podría apoyar a los inversores en la toma de decisiones informadas.

El desarrollo reciente de modelos financieros avanzados, como FinGPT, XuanYuan y DISC-FinLLM, se ha enfocado principalmente en el preentrenamiento o ajuste supervisado, aunque aún se subutilizan métodos como el Retrieval Augmented Generation (RAG) para integrar conocimiento financiero especializado (Li et al., 2024). La construcción de estos FinLLMs depende de la recopilación de grandes volúmenes de datos multimodales y del uso de técnicas avanzadas que mejoran la precisión y la fiabilidad de las respuestas generadas, al tiempo que abordan cuestiones de privacidad en los datos (Lee et al., 2025). En esta línea, Onan and Dursun (2024) proponen la adopción de sistemas RAG para lograr análisis más detallados y robustos en mercados complejos.

Varios estudios recientes han explorado aplicaciones innovadoras de los LLMs en finanzas. Ma et al. (2024) demuestran que modelos como ChatGPT pueden predecir primas de riesgo de acciones a partir de titulares de noticias, superando métodos tradicionales de análisis de sentimiento. Shao et al. (2024) presentan el modelo HD-SURDLM, que integra análisis de sentimiento avanzado y modelado financiero dinámico, mejorando el pronóstico en horizontes de corto y largo plazo. Así mismo, Kang et al. (2024) examinan el impacto del sentimiento del mercado, modelado con ChatGPT, sobre la predicción de riesgo en el mercado de Bitcoin, utilizando simulación de Monte Carlo para calcular el valor en riesgo (VaR). En gestión de carteras, Ko and Lee (2024) encuentran que ChatGPT produce selecciones de portafolio mejor diversificadas y de mayor rendimiento que las aleatorias, evidenciando su utilidad como asistente para gestores de inversiones.

5.4.1 Limitaciones y Oportunidades de los FinLLMs

A pesar del progreso alcanzado por los FinLLMs, persisten áreas relevantes para su perfeccionamiento, especialmente en tareas propias del dominio como la detección de causalidad, el razonamiento numérico y la identificación de eventos corporativos relevantes (Lee et al., 2025). Mejoras en estos campos pueden influir significativamente en la toma de decisiones basada en datos financieros. Para adaptar los LLMs generales a las tareas específicas del dominio, se ha recurrido al ajuste fino eficiente (PEFT) y a la ingeniería de prompts, elevando su desempeño en la gestión de grandes

volúmenes de información financiera (R. Mukherjee et al., 2022). No obstante, la escasez de conjuntos de datos especializados en finanzas sigue representando un desafío central para el desarrollo de estos modelos. Por último, la integración de reportes generados por resumen abstractivo, en conjunto con el uso de LLMs, permite ampliar tanto el alcance como la calidad de las soluciones aplicadas al análisis de mercados financieros, tal como se propone en el presente trabajo.

En conjunto, los conceptos revisados en este marco permiten sustentar técnicamente la solución propuesta en este trabajo de grado. La integración de modelos de lenguaje, técnicas de resumen abstractivo, evaluación automática y métodos de ajuste eficiente como LoRA y destilación de conocimiento, no solo responde a los desafíos actuales del procesamiento de información en el sector financiero, sino que habilita una solución profesionalizable, escalable y sostenible. Esta fundamentación teórica respalda el desarrollo de una herramienta orientada al usuario final, con capacidad de ser implementada en entornos reales para apoyar la toma de decisiones de inversión, lo que reafirma la pertinencia del proyecto dentro de una maestría profesionalizante en ciencias de los datos y la analítica.

6 Diseño Metodológico

El objetivo fundamental de este proyecto reside en el desarrollo de una herramienta robusta y accesible capaz de generar reportes resumidos personalizados para inversores individuales, utilizando información obtenida de análisis financieros y de trading proporcionados por expertos a través de videos en plataformas como YouTube. Esta funcionalidad se orienta a brindar apoyo oportuno en la toma de decisiones de inversión.

La metodología implementada se fundamenta en una aproximación híbrida, integrando el proceso de ciencia de datos en equipo (TDSP) de Microsoft para la estructuración general (microsoft, 2020) y la gestión iterativa del proyecto, junto con el ciclo de vida de un proyecto de Inteligencia Artificial Generativa (AIGC) propuesto por (Bandi and Kagitha (2024)), el cual guía las fases específicas inherentes a la generación de contenido mediante modelos de lenguaje. El TDSP, con sus fases de comprensión del negocio, adquisición y comprensión de datos, modelado, implementación y aceptación del cliente, proporciona un marco ágil y colaborativo. Ver Figura 1 tomada de microsoft (2020). Paralelamente, el ciclo AIGC, compuesto por la identificación del caso de negocio, la definición y obtención de datos, la selección del modelo LLM, su adaptación y, finalmente, la implementación, se aplica para optimizar la generación de los reportes tipo resumen de las transcripciones de los videos. Ver Figura 2 tomada de Bandi and Kagitha (2024).

Fases del Ciclo de Vida del TDSP: El ciclo de vida del TDSP se compone de cinco fases principales que se ejecutan de manera iterativa: (1) Conocimiento del negocio, donde se comprenden los objetivos y requisitos del proyecto; (2) Adquisición y comprensión de los datos, que implica la recopilación y análisis de datos relevantes; (3) Modelado, donde se desarrollan modelos predictivos utilizando técnicas de machine learning; (4) Implementación, que consiste en desplegar los modelos en un entorno de producción; y (5) Aceptación del cliente, donde se validan los resultados con los stakeholders para asegurar su satisfacción. Ver Figura 1 tomada de microsoft (2020).

Según Bandi and Kagitha (2024) un **AIGC** se estructura en cinco fases principales: primero, se identifica el caso de negocio y se define el alcance, estableciendo el tipo de contenido a generar; segundo, se definen y obtienen los datos necesarios para personalizar el modelo, aprovechando conjuntos de datos masivos; tercero, se selecciona el modelo LLM adecuado, evaluando opciones y equilibrando costos y rendimiento; cuarto, se adapta el LLM al caso de uso mediante ingeniería de prompts, ajuste fino y/o técnicas de aprendizaje por refuerzo con retroalimentación humana (RLHF); y finalmente, se implementa el sistema y modelo mediante plataformas en la nube para simplificar la gestión de recursos.

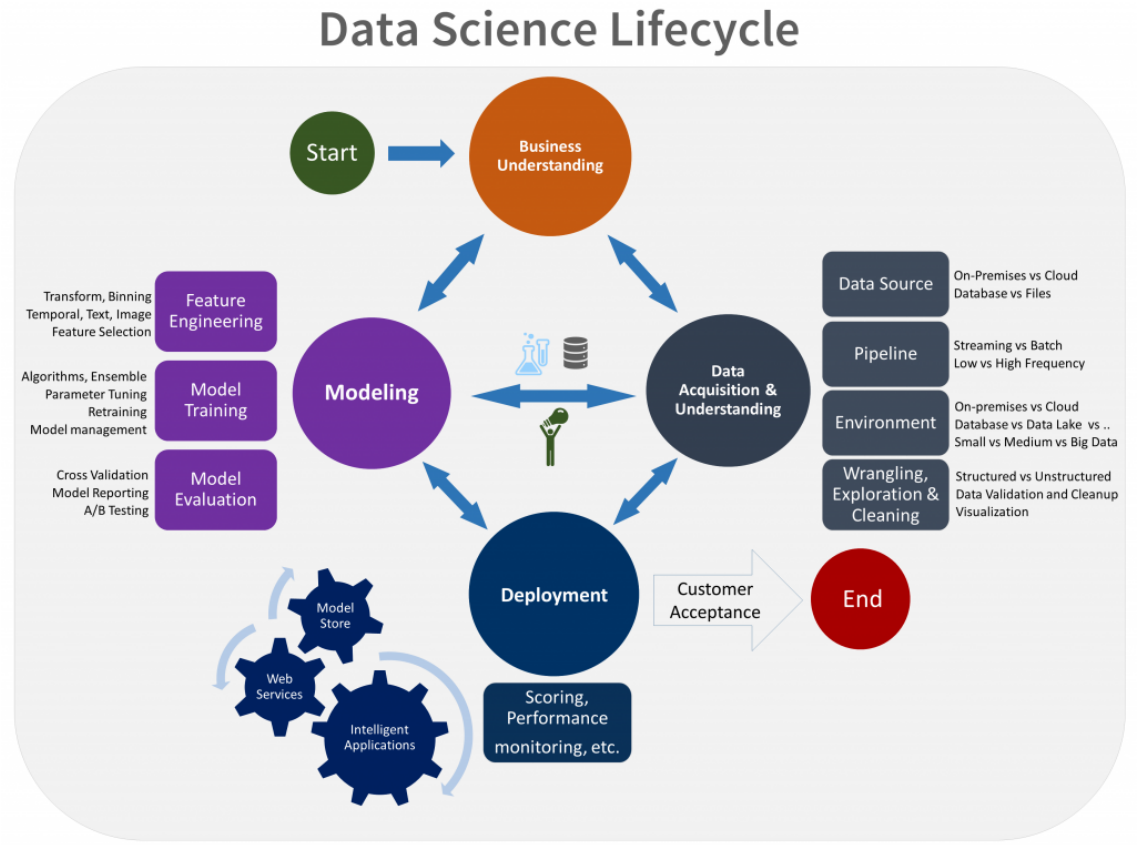

Figure 1: TDSP Ciclo de vida de ciencia de datos.

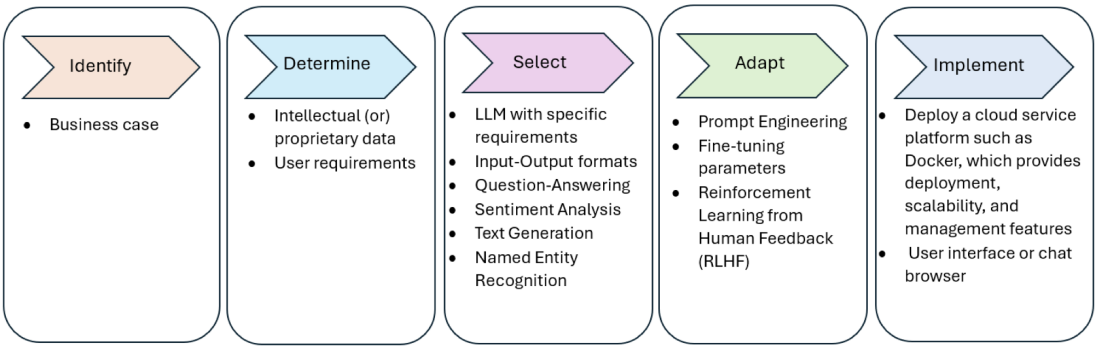

Figure 2: Fases del ciclo de vida de un proyecto de IA generativa.

6.1 Etapas metodológicas

6.1.1 Comprensión del negocio

Problema identificado: El mercado financiero contemporáneo se caracteriza por su alta complejidad y un volumen masivo de datos estructurados y no estructurados que se generan diariamente. Los inversores individuales se enfrentan a una sobrecarga de información y limitaciones en la capacidad para procesar y evaluar estos datos, lo que a menudo resulta en decisiones basadas en información incompleta o desactualizada. La interpretación de datos provenientes de múltiples fuentes (análisis fundamental, técnico, sentimiento en redes sociales) es compleja, y los métodos tradicionales de análisis financiero son insuficientes. Además, existe una barrera de acceso a herramientas de análisis avanzadas para inversores individuales, a diferencia de las grandes instituciones. La información valiosa a menudo reside en fuentes no estructuradas, como las transcripciones de videos, cuya extracción de insights requiere técnicas avanzadas de PLN no siempre accesibles de forma gratuita o a un costo razonable. Finalmente, el alto costo computacional de los Modelos de Lenguaje Grandes (LLMs) de última generación limita su implementación generalizada en aplicaciones reales a no ser que se pague por su uso a traves de servicios en la nube.

Meta del proyecto: La meta principal es asistir a los inversores individuales mediante la generación de reportes eficientes y accesibles. Esto se logra mediante el ajuste fino de un Modelo de Lenguaje Pequeño (SLM) para extraer y resumir información clave de transcripciones de videos de análisis financiero de expertos en YouTube. El objetivo es democratizar y facilitar el acceso a información financiera experta, mejorando la capacidad de los usuarios para tomar decisiones fundamentadas con una inversión mínima de tiempo y recursos.

6.1.2 Adquisición de datos

Proceso de ingesta: Se estableció un proceso de ingesta automática diaria de transcripciones y metadatos de videos mediante una función serverless en Azure (Azure Function App).

Fuentes de datos Las transcripciones se obtienen de videos publicados por seis canales especializados de YouTube (ARENA ALFA, Bolsas hoy — Invierte y Crece, Esteban Pérez, USACRYPTONOTICIAS, Bitcoin hoy, Bolsas hoy — Esteban Pérez), que realizan análisis fundamentales y técnicos de activos financieros y criptomonedas.

Almacenamiento: Los datos ingestados se procesan y almacenan en un datalake en Azure, pasando por una capa landing a una capa raw y finalmente a una capa curated.

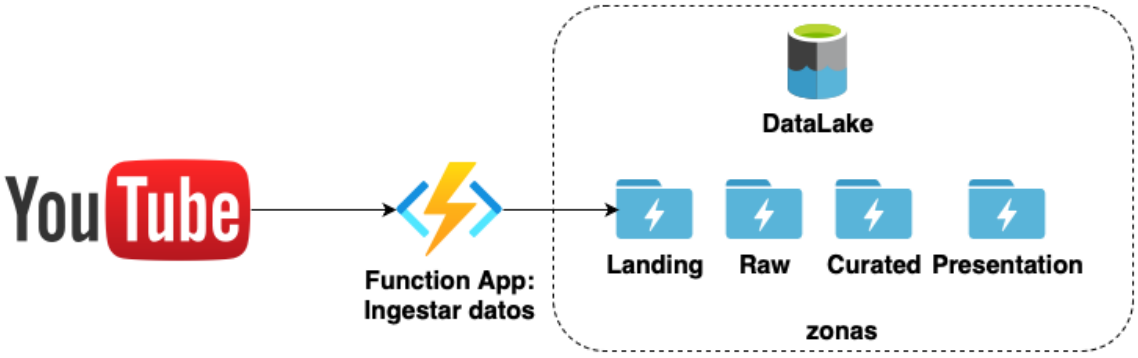

Figure 3: Ingesta datos canales de youtube a Datalake.

El código de la implementación de la azure function para ingestar los datos en el datalake, en un grupo de recursos en la nube de azure, se encuentra en el repositorio del Anexo A.1.

6.1.3 Preparación de los datos

Proceso de ELT (Extracción-Carga-Transformación):

Capa raw: Los datos se limpian, se eliminan duplicados basándose en el `video_id`, se completan valores nulos de `duration`, y se descartan columnas irrelevantes, como `keywords` o `description`, añadiendo fechas clave. Los datos se almacenan en formato Parquet, particionados por año.

Capa curated: Se valida la disponibilidad de registros, se limpian y transforman los textos (eliminando emoticones, normalizando y creando nuevos campos combinados de título y subtítulos), y se seleccionan y renombran columnas relevantes para el análisis.

Creación del conjunto de entrenamiento: Se generó un conjunto de datos sintético "from scratch" para entrenamiento y prueba. Esto se hizo utilizando el modelo fundacional pre-entrenado gpt4.1 (versión `gpt-4_1-2025-04-14`) a través de la API de OpenAI. El objetivo fue crear resúmenes en un formato de reporte específico, aprovechando el conocimiento de este modelo de billones de parámetros. El conjunto de datos resultante contiene 2225 resúmenes, con un costo de $44,107 USD para procesar aproximadamente 15,8 millones de tokens de entrada y generar 1,77 millones de tokens. En Figura 4 se muestra el flujo de trabajo para la creación del conjunto de datos de entrenamiento. El proceso de creación del conjunto de datos se puede ver en el notebook del Anexo A.2 `notebooks/create_summaries_v2.ipynb` y el conjunto de datos creado está en el Anexo A.3 con un conjunto de datos de entrenamiento y evaluación.

En Pieper et al. (2024) se demuestra el potencial de utilizar LLMs para generar conjuntos de datos, mejorando el rendimiento de modelos más pequeños en tareas complejas y ofreciendo un método rentable para su ajuste fino. Esta elección metodológica se apoya en la discusión previa sobre el impacto de la generación de

datos y la destilación de conocimiento para la transferencia de capacidades desde modelos fundacionales a modelos de menor cantidad de parametros.

Estructura del conjunto de datos: El conjunto de datos se dividió en 2004 registros para entrenamiento y 221 para evaluación. Cada registro corresponde a un video ingestado y contiene las siguientes 15 columnas:

channel_name: Nombre del canal de YouTube donde se publicó el video.

video_id: Identificador único de cada video, código alfanumérico que aparece en la URL del video.

source: URL del video en "YouTube".

publish_date: Fecha en la que el video fue publicado originalmente. A algunos registros (115) les falta esta información.

duration: Duración del video en segundos, almacenada como un número decimal.

last_update_date: Fecha de la última actualización de los datos del video en este conjunto de datos.

title: Título del video como aparece en YouTube.

text: Transcripción del video.

year: Año en el que se publicó el video, extraído de la fecha de publicación.

month: Mes en el que se publicó el video, extraído de la fecha de publicación.

number_of_tokens: Conteo de tokens en la transcripción completa del video.

prompt: Prompt aplicado a GPT-4.1 para resumir y extraer los términos de la transcripción del video.

summary: Resumen de la transcripción del video generado con modelo GPT-4.1.

key_terms: Palabras clave que representan activos financieros y criptoactivos extraídos por el modelo GPT-4.1 en la transcripción del video.

index_level_0: Columna de índice que indica la posición de la fila en el conjunto de datos original.

El conjunto de datos usa aproximadamente 260.9 KB de memoria. Para realizar un análisis de costos e implementación se crea la columna summary_tokens y se transforma la columna duration de segundos a minutos.

En la Tabla 1 se puede observar un ejemplo de una transcripción extraída de un video de YouTube, de uno de los canales seleccionados. Con el respectivo resumen y la extracción de activos mencionados usando el modelo gpt4.1.

Texto	Resumen	Activos (palabras clave)
qué hará bitcóin hoy 02.01.2024 08:00 análisis técnico `btc-eth` esteban perez trader. saludos cordiales y bienvenido como vemos en la imagen algo que no logró llegar a ocurrir en los últimos días en las últimas dos semanas del año 2023 que los futuros de bitcóin en el mercado `cme` obtuvieron sus objetivos de corto plazo y también de mediano plazo vemos que en estas horas en cuanto ha abierto la cotización en este año 2024 en los futuros ya lo ha logrado en este momento podemos ver que son las 8 de la mañana de Madrid las 7 de horario gmt el horario universal y damos comienzo como cada jornada a partir de ahora cada día hasta el viernes con este primer análisis donde incluiremos niveles intr diarios [música] comenzamos y lo hacemos de la mano como siempre mirando antes que nada los fundamentales el calendario macroeconómico para la jornada de hoy que lo encuentras en la plataforma de xtv desde estados unidos el índice pmi para la industria del mes de diciembre más tarde en el canal de la bolsa hoy que tienes abajo el enlace por si no lo conoces ya analizaremos con mayor profundidad el calendario para esta semana también para este mes y miraremos todos los aspectos no solo de bitcoin y ethereum que están incluidos sino de todos los activos que vamos viendo en la bolsa y ahora nos concentramos en bitcoin estamos en el contado en este momento si viste el análisis de ayer domingo donde comenzamos con los primeros niveles de refugio de mediano plazo desde la apertura del año 42331 que también lo tenemos en cuenta la línea amarilla este era nuestro eje central de operaciones para esta semana todo lo que fuera estar por encima de ese nivel también ahora vemos cómo ha quedado la situación en 4 horas nuestros objetivos semanales los teníamos el primero en los 43385 aquí 44100 y a partir de ahí tenemos el siguiente objetivo colocado en los 46,560 primer refugio de mediano plazo calculado para este año vamos a mirar la serie de 4 horas porque efectivamente en ese análisis estábamos hablando de que la probabilidad que se generara una nueva fase alcista y que hubiera cambio de niveles era bastante alta ya que no hacía el precio sino romper la zona de máximos previos decrecientes	**Introducción** El análisis técnico presentado por Esteban Pérez, trader, se centra en las perspectivas de Bitcoin y Ethereum al inicio del año 2024. El informe detalla los movimientos recientes en los mercados de futuros y contado, destacando niveles clave y posibles escenarios para ambos activos. Se hace hincapié en la importancia de los niveles intradiarios y se ofrece un análisis detallado de las tendencias actuales y futuras. **Puntos más importantes** - **Futuros de Bitcoin en CME**: Al inicio de 2024, los futuros de Bitcoin han alcanzado sus objetivos de corto y mediano plazo, algo que no se logró en las últimas semanas de 2023. - **Niveles clave de Bitcoin**: - Eje central de operaciones para la semana: 42,331. - Objetivos semanales: 43,385, 44,100, y 46,385. - Posible formación de un patrón de "hombro cabeza hombro" en la acción del precio. - Triple máximo roto en la zona de 44,444. - **Análisis de Ethereum**: - Punto de partida: 2,282.10. - Objetivos semanales: 2,443 y 2,515. - Fase alcista tanto en semanal como en diario. - **Eventos macroeconómicos**: Se menciona la importancia del índice PMI de Estados Unidos para la industria de diciembre, que podría influir en los mercados. - **Tendencias y escenarios**: - Bitcoin muestra una nueva fase alcista con niveles clave de soporte y resistencia. - Ethereum también se encuentra en una fase alcista, con niveles intradiarios importantes para el trading. **Conclusión** El análisis técnico de Esteban Pérez sugiere un inicio de año prometedor para Bitcoin y Ethereum, con ambos activos mostrando tendencias alcistas y alcanzando niveles clave. Los traders deben estar atentos a los niveles intradiarios y a los eventos macro económicos que podrían influir en el mercado. La atención se centra en cómo se desarrollarán estas tendencias en los próximos días y semanas. **Activos mencionados para invertir** - Bitcoin (BTC) - Ethereum (ETH)	`['Bitcoin (BTC)', 'Ethereum (ETH)']`

Texto	Resumen	Activos (palabras clave)
esta línea azul se rompió finalmente estuvimos viendo que este escenario se estaba dando con mayor probabilidad que tenía un primer objetivo en 44,100 44,444 y ahora 46385 es el objetivo de toda esta subida fíjate que está muy cercano apenas a 200 menos de 00 del refugio de mediano plazo el primero ya se ha conseguido para esta semana y de momento el segundo está a tiro de piedra antes de continuar precisamente con todos estos de muy corto plazo también trabajábamos con la idea de que se generara un triple máximo pero esto lo veníamos abordando desde las semanas anteriores del año pasado de que el precio quiera subir dejara un nuevo máximo dejar una trampa alcista y después pudiésemos ver el siguiente movimiento bueno de momento ya tenemos triple máximo que además ha roto el anterior la zona de los 44000 400 44700 4469 en concreto bueno es esa zona ya se ha roto también la de los 44,444 abrimos un poco el gráfico y se ve bastante bien la figura que se está formando aquí tenemos la primera la segunda la tercera posibilidad de que se vea como una especie de hombro cabeza hombro en la acción del precio que es lo que tratamos aquí es una situación de triples máximos bueno pues así está en la serie diaria han roto con bastante fuerza los 44100 en la serie de 4 horas entonces este escenario quedó ad nunca se llegó a realizar el escenario donde se pudieran disparar las ventas en corto y distribución de ganancias no se llegó a validar como bien sabes ahora tenemos ya esta nueva fase alcista con los siguientes niveles 42503 vemos cómo fue funcionando 4287 en cuanto lo tocó una reacción de vuelta a la apertura del año desde aquí a partir de los 42000 43245 que estás viendo que cambia de color también con una situación de eventos que se han ido produciendo máximos decrecientes que se rompió referencias de pivot anteriores que vinieran decreciendo también se rompió a partir de los 43245 por lo tanto		

Table 1: Resumen y extracción de términos transcripción video.

Consideraciones del lenguaje hablado: Las transcripciones presentan desafíos adicionales debido a la naturaleza menos estructurada del lenguaje hablado, in-

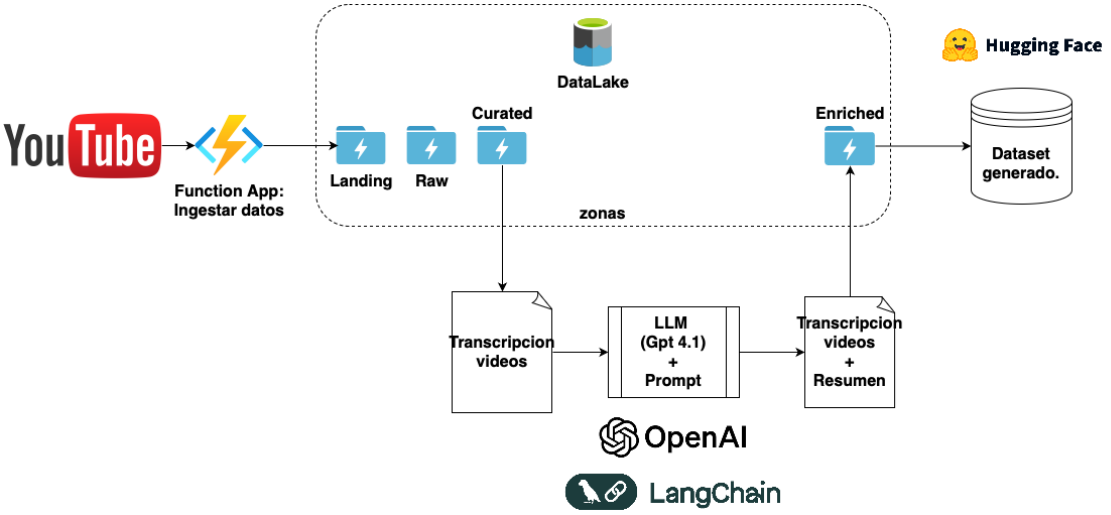

Figure 4: Creación del conjunto de datos de entrenamiento y evaluación.

cluyendo frases incompletas, repeticiones y cambios abruptos de tema, así como la presencia de jerga técnica financiera y referencias a imágenes que se están mostrando en el video como contexto adicional. La identificación y resumen de puntos clave es más desafiante que en textos escritos tradicionales.

6.1.4 Modelado

En la Figura 5 se puede observar el flujo de trabajo general que se siguió en la etapa de modelado para el ajuste fino del modelo SLM.

Arquitectura y modelo base: Se utilizaron Modelos de Lenguaje Pequeños (SLMs), incluyendo versiones de Llama 3.2 (1B y 3B parámetros) y Llama 3.1 (8B parámetros) como modelos base.

Técnica de ajuste fino: Se aplicó la técnica de instruction fine-tuning para adaptar los modelos a la tarea de generar resúmenes con un formato de reporte específico. Para optimizar los recursos computacionales y la eficiencia, se empleó el método de **Ajuste Eficiente de Parámetros (PEFT)**, específicamente **LoRA (Low-Rank Adaptation)**. LoRA permite adaptar modelos grandes congelando sus pesos originales e introduciendo matrices de bajo rango, lo que reduce significativamente los parámetros entrenables y los requisitos de hardware sin aumentar la latencia de inferencia.

Preparación del dataset para fine-tuning: Las transcripciones de los videos y sus resúmenes se transformaron en un formato de pares instrucción/entrada/respuesta, con la adición de tokens especiales, para alinear el modelo y que siga instrucciones de manera precisa.

Librerías y recursos: Se utilizaron las librerías litgpt y unsloth para el proceso de entrenamiento. unsloth fue fundamental por su eficiencia en el fine-tuning de modelos de lenguaje grandes, especialmente para manejar ventanas de contexto largas (hasta 8192 o 16384 tokens en algunos experimentos) y mitigar errores de falta de memoria (OOM). Para el fine-tuning de Llama 3.1 de 8B parámetros se utilizó una GPU NVIDIA A100-SXM4 con 40 GB de RAM.

Configuración de hiperparámetros: Los parámetros de entrenamiento incluyen el learning rate, batch size, epochs, tamaño de bloque, y los valores de lora_r y alpha. Se exploraron precisiones como bf16-true y métodos de cuantización como bnb.nf4 para reducir el consumo de memoria GPU.

Para el entrenamiento, se cargan modelos base preentrenados. En la generación de resúmenes tipo reporte, es fundamental considerar que los textos de entrada (transcripciones) requieren una longitud de secuencia máxima (`max_seq_length`), determinada en nuestro caso en 42,638 tokens con base en los datos de entrenamiento. Este parámetro define la mayor cantidad de tokens que el modelo puede procesar durante el entrenamiento o la evaluación; secuencias más largas exigen un mayor consumo de memoria y cómputo. Esto limita el uso de modelos con mayor contexto y/o con mayor número de parámetros, como Llama 3.1 de 8 billones de parámetros.

Otro factor relevante es la precisión de los cálculos numéricos: 32 bits (`32-true`), 16 bits bfloat (`bf16-true`), o precisión mixta (`bf16-mixed`). El uso de menor precisión reduce el consumo de memoria y puede acelerar el entrenamiento, aunque eventualmente puede impactar la estabilidad o calidad del modelo.

La cuantización (`Quantize`) permite reducir el tamaño del modelo utilizando formatos más ligeros (por ejemplo, 8 bits, 4 bits con `bitsandbytes`, `bnb.nf4`, etc.), lo que resulta especialmente útil en GPUs con recursos limitados.

Para aprovechar secuencias de entrada largas, es importante tener en cuenta el tamaño del modelo y ajustar la precisión numérica y el método de cuantización al cargar el modelo, optimizando así el uso de memoria GPU. Además, debe ajustarse el tamaño de `batch` y evitar el cálculo del error en el conjunto de validación durante el entrenamiento, ya que, en pruebas, se evidenció que esto genera un alto consumo de memoria que puede resultar en errores de *out-of-memory* (OOM).

Las épocas (`epochs`) definen el número de veces que el conjunto de entrenamiento es recorrido en su totalidad.

El tamaño de `global_batch` corresponde al número total de muestras procesadas por iteración, sumando todos los dispositivos y acumulaciones. Este valor controla la estabilidad del entrenamiento y el consumo de memoria.

La variable `micro_batch_size` define la cantidad de muestras procesadas antes de acumular gradientes, es decir, por dispositivo (por ejemplo, en una sola GPU). Utilizar micro-batches más pequeños permite entrenar modelos grandes en GPUs de capacidad limitada, mediante la acumulación de gradientes.

En resumen:

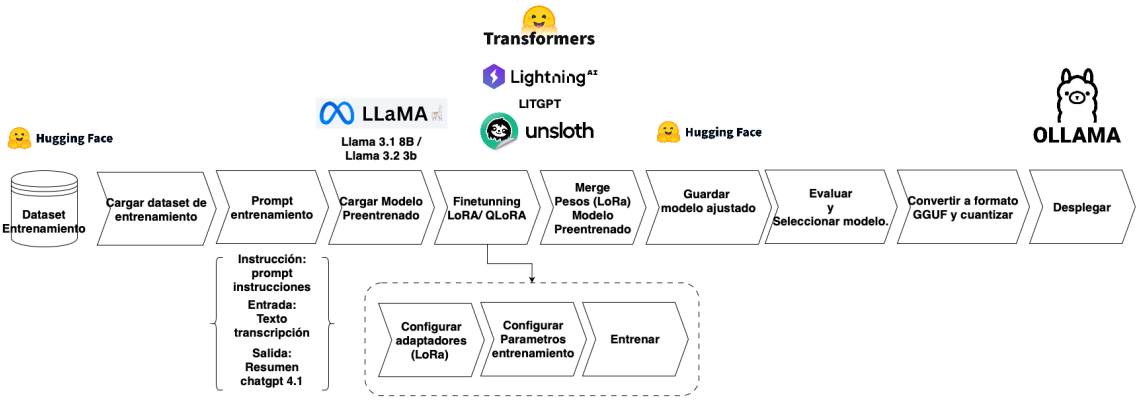

Figure 5: Flujo de trabajo general de entrenamiento, evaluación y despliegue.

- `max_seq_length` controla la memoria y la longitud del contexto.
- Los parámetros de precisión y `quantize` ajustan el consumo de memoria (a costa de precisión/calidad).
- `global_batch_size` y `micro_batch_size` deben adaptarse para equilibrar consumo de memoria y estabilidad durante el entrenamiento.
- `epochs` permite regular el tiempo de entrenamiento y el riesgo de sobreajuste/subajuste.

Aunque las métricas `val_loss` y `val_ppl` son útiles para monitorizar el sobreajuste y el progreso del modelo, su cálculo incrementa significativamente el uso de memoria de la GPU. Por ello, para modelos con secuencias de entrada largas y más de 3 billones de parámetros —como en nuestro caso de uso—, se recomienda evitar la validación durante el entrenamiento.

Monitoreo del entrenamiento: Para el seguimiento de los experimentos y la evaluación de modelos, se emplearon Weights & Biases (W&B) y MLflow, respectivamente.

6.1.5 Evaluación

La evaluación de los modelos ajustados se llevó a cabo siguiendo un flujo de trabajo estructurado en la Figura 6, donde las respuestas generadas por los modelos fueron comparadas con un resumen de referencia creado previamente utilizando el modelo fundacional GPT-4.1, ver sección 6.1.3. Este conjunto de prueba no contiene resúmenes generados por humanos. Para este proceso, se dividió el conjunto de datos en un 90% para entrenamiento (2004 transcripciones) y un 10% para prueba (221 transcripciones), asegurando un balance por canal.

Distancia de Coseno: Se calculó la distancia de coseno entre los embeddings, generados con el modelo `text-embedding-3-large` , de la respuesta del modelo ajustado y la respuesta de referencia (generada por GPT-4.1). Una menor distancia indica una mayor similitud semántica y, por ende, una mejor respuesta.

Evaluación con LLM-as-a-Judge : El proceso de evaluación de los resúmenes se complementó con el uso de Modelo de Lenguaje Grande (LLM) como evaluador, una práctica que ha ganado prevalencia debido a la alta correlación de sus puntuaciones con las asignadas por evaluadores humanos, superando en muchos casos a las métricas automáticas tradicionales. Para este proyecto, se empleó un modelo fundacional (GPT-4.1) en un rol de evaluador.

La metodología incluyó la definición de un prompt evaluador que instruía al modelo para generar dos tipos de puntuaciones: un score simple en un rango de 0 a 10 y un score basado en criterios específicos. Los criterios de evaluación fueron:

- **Fidelidad (Faithfulness):** asegurar que no se inventara información.
- **Relevancia (Relevance):** verificar que el resumen se centrara en los puntos importantes.
- **Concisión (Conciseness):** evaluar la brevedad y la ausencia de relleno.
- **Coherencia (Coherence):** determinar si el resumen estaba bien estructurado y era de fácil lectura.

Este enfoque, donde el LLM actúa como evaluador, se planteó como un elemento central para la validación metodológica y el análisis de resultados, ofreciendo una alternativa eficiente y escalable a la revisión manual.

6.1.6 Despliegue

Entorno de despliegue: Para la prueba de concepto (PoC), se implementó la aplicación localmente en una máquina con Chip M3 Pro (CPU de 12 núcleos, GPU de 18 núcleos y 36 GB de RAM unificada).

Interfaz de usuario: Se desarrolló una interfaz gráfica de usuario utilizando la librería Streamlit, que permite desplegar aplicaciones web con pocas líneas de código. Para detalles del código ver el repositorio en el Anexo A.2

Modelo en despliegue: Los modelos seleccionados, previamente guardados en formato GGUF en HuggingFace, se descargan localmente y se despliegan utilizando la aplicación Ollama, facilitando la ejecución de LLMs.

Funcionalidades del prototipo: La interfaz gráfica dispone de tres secciones principales, ver Figura 7. :

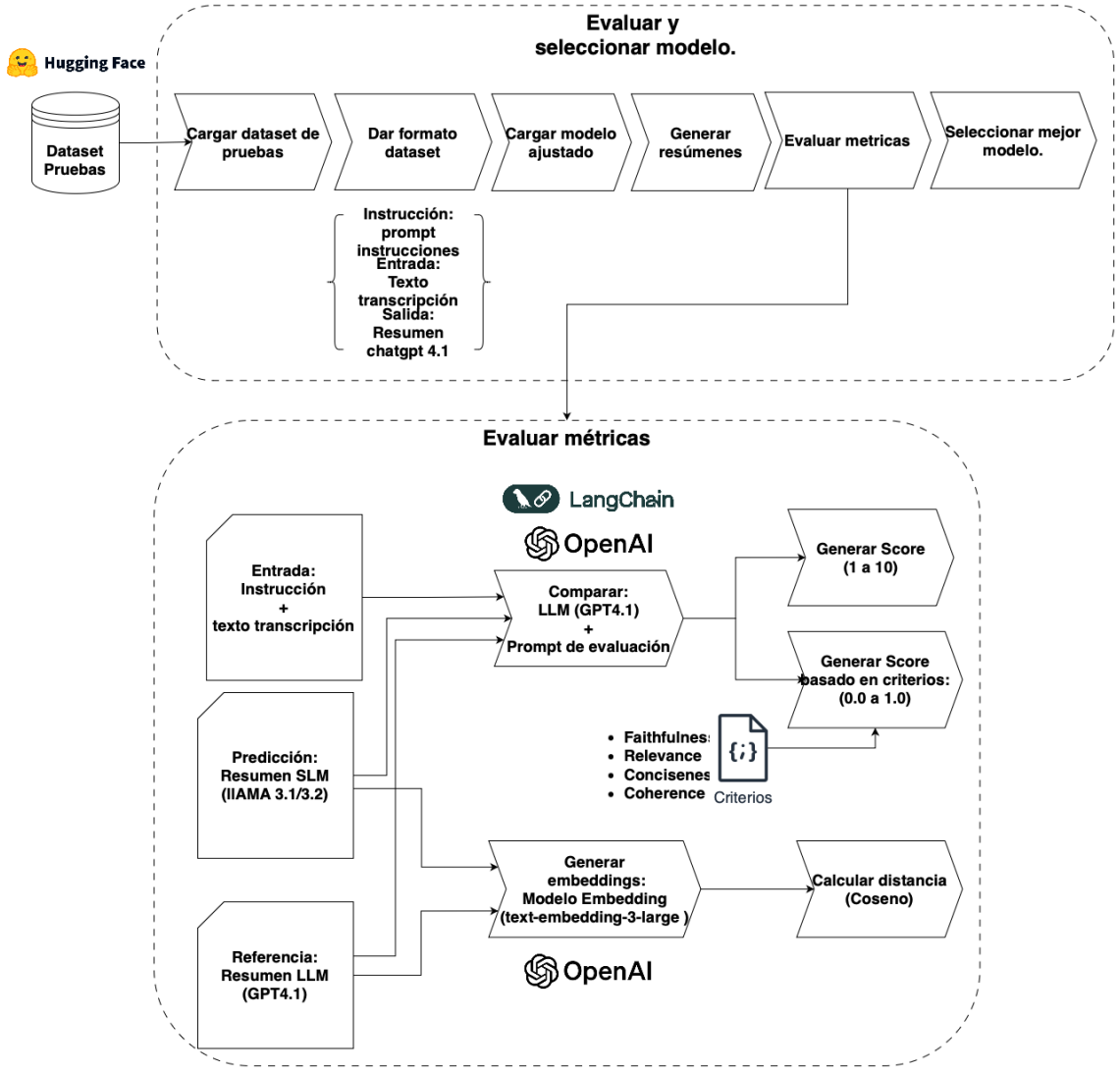

Figure 6: Flujo de evaluación de los modelos.

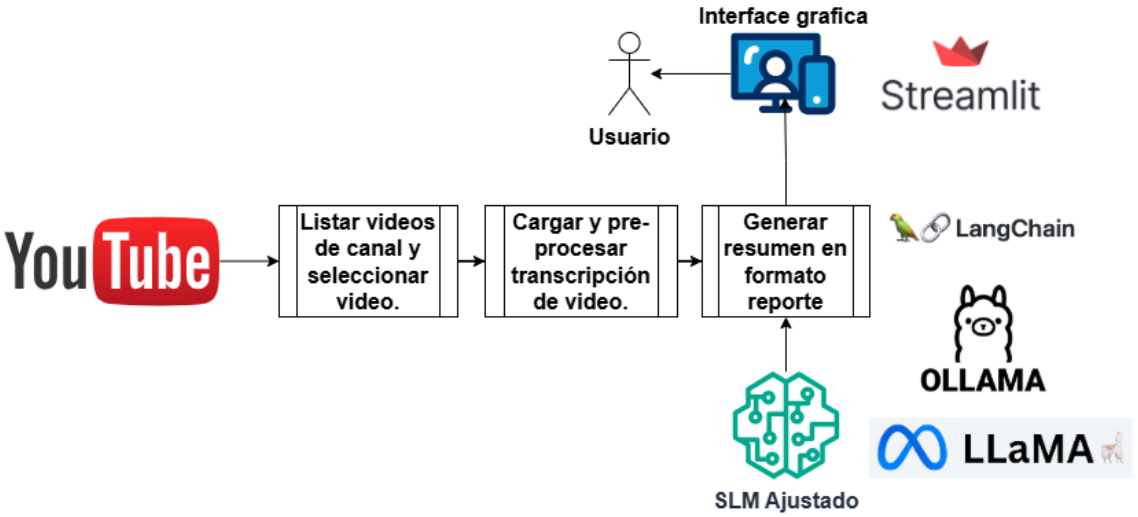

Figure 7: Despliegue del modelo.

Generación de reportes resumidos de videos: Permite seleccionar o introducir la URL de un video, así como el modelo y el prompt para generar el reporte.

Reporte de métricas: Permite comparar el rendimiento de los diferentes modelos en el conjunto de evaluación.

Visualización detallada de evaluaciones: Muestra en detalle las evaluaciones de los LLMs en el proceso de generar los scores simples y criteriales.

7 Resultados

Esta sección detalla los resultados del análisis exploratorio de los datos, los resultados de la experimentación y la evaluación de los modelos de lenguaje ajustados, presentando los hallazgos de forma cuantitativa y cualitativa, junto con una reflexión sobre su cumplimiento de los objetivos y las limitaciones identificadas.

7.1 Análisis exploratorio del conjunto de datos (EDA)

En la Tabla 2, observamos estadísticas básicas del conjunto de datos creado de transcripciones de videos los canales seleccionados. Esto es valioso para estimar el costo del procesamiento por tokens y ver las ventanas de contexto para el LLM y/o SLM por video.

	Numero tokens transcripciones.	**Numero de tokens resúmenes generados.**	**Duración de videos en minutos.**
Conteo datos.	2.225	2.225	2.225
promedio	7.119	796	42,8
desviación estándar	6.954	235	43,5
valor mínimo	34	172	0,0
cuartil inferior (25%)	2.280	632	12,9
mediana	3.408	799	19,2
cuartil superior (75%)	11.767	957	71,6
Valor máximo	42.638	1.910	349,6

Table 2: Estadística básica de transcripciones videos YouTube

El tiempo total de los videos de los canales seleccionados por año nos permite ver aproximadamente al año cuanto tiempo debe invertir una persona en ver los análisis de expertos para obtener información relevante del mercado y de posibles oportunidades de inversión, para el año 2024 fueron 497 horas y para el primer semestre de 2025 se requieren 204 horas.

En la Figura 10 observamos que el canal que más publica videos en promedio es el canal USA CRIPTONOTICIAS, seguido del canal Esteban Perez, en los últimos años y lo que va del primer trimestre de 2025. Pero debemos tener en cuenta que el año 0 corresponde a las transcripciones sin fecha de publicación (Que se distribuyen en estos años).

En la Figura 11 podemos ver el tiempo promedio de duración de los videos por canal y año, donde podemos observar que el canal USACRYPTONOTICIAS tiene el promedio más alto de duración de los videos. Esto es importante ya que nos puede ayudar a tomar decisiones en el despliegue e implementación de la herramienta. Para evaluar temas de costos ya sea por consumo del API de los modelos pagos y/o costo computacional debemos llevar este mismo análisis a los tokens (calculados

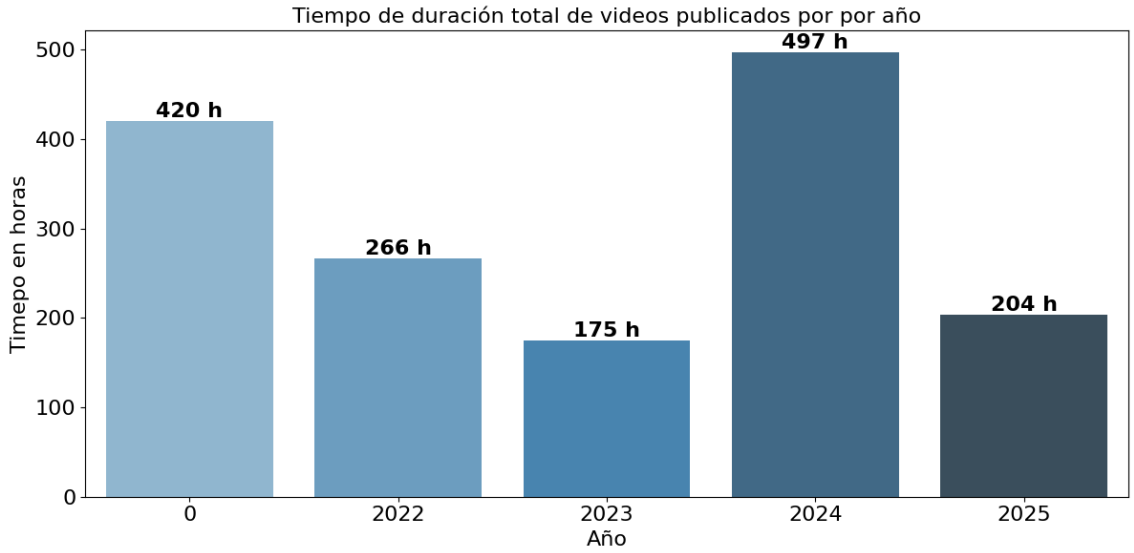

Figure 8: Duración total de videos por año.

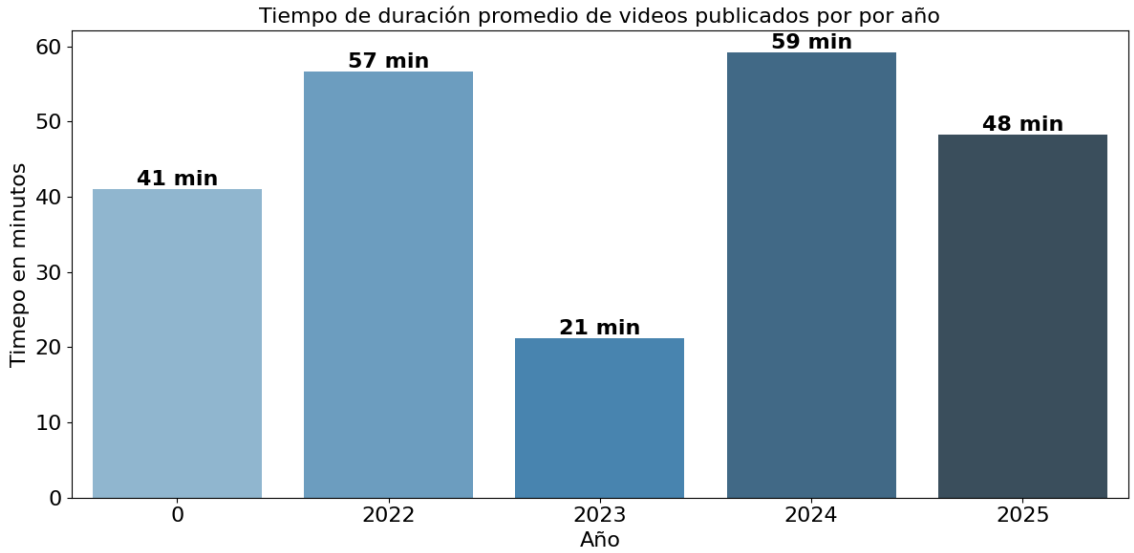

Figure 9: Duración promedio de videos por año.

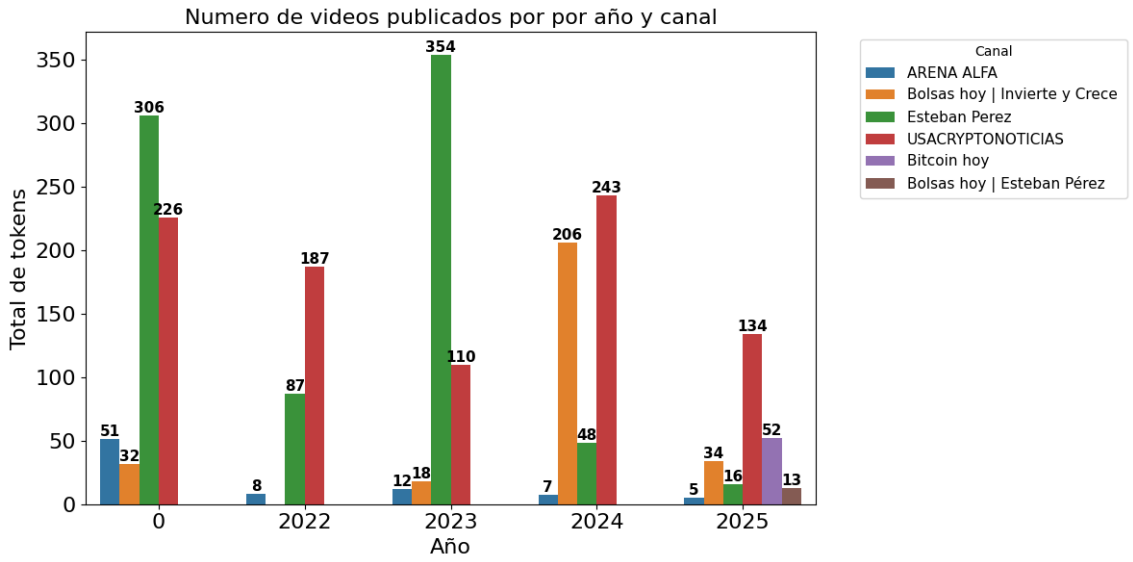

Figure 10: Videos publicados por canal y año.

con la libreria tiktoken y para el modelo gpt4o) de las transcripciones de los videos, en la Figura 12 podemos observar el número de tokens totales y promedio de las transcripciones de los videos por año y canal, donde observamos que los tokens medios de las transcripciones del canal USACRIPTOMONEDAS son mucho mayores que los de los demás canales.

En la Figura 13 podemos analizar el año 2024 y lo que va del 2025, para observar el consumo de tokens mínimo, máximo y medio de las transcripciones de los videos por mes. Este análisis es importante para evaluar costos mensuales medios y también poder evaluar recursos en el despliegue de la aplicación. Recordemos que si estamos usando por ejemplo el API de openIA el costo de la generación de texto esta dado por los tokens de entrada más los tokens generados (para el caso específico del modelo gpt 4.1.)

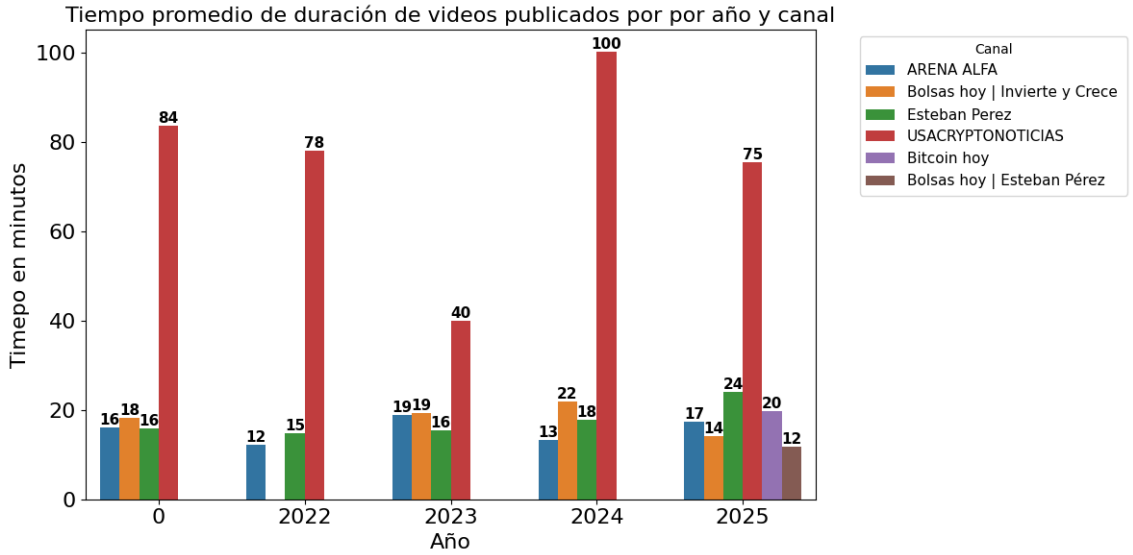

Figure 11: Duración promedio de videos publicados por año y canal.

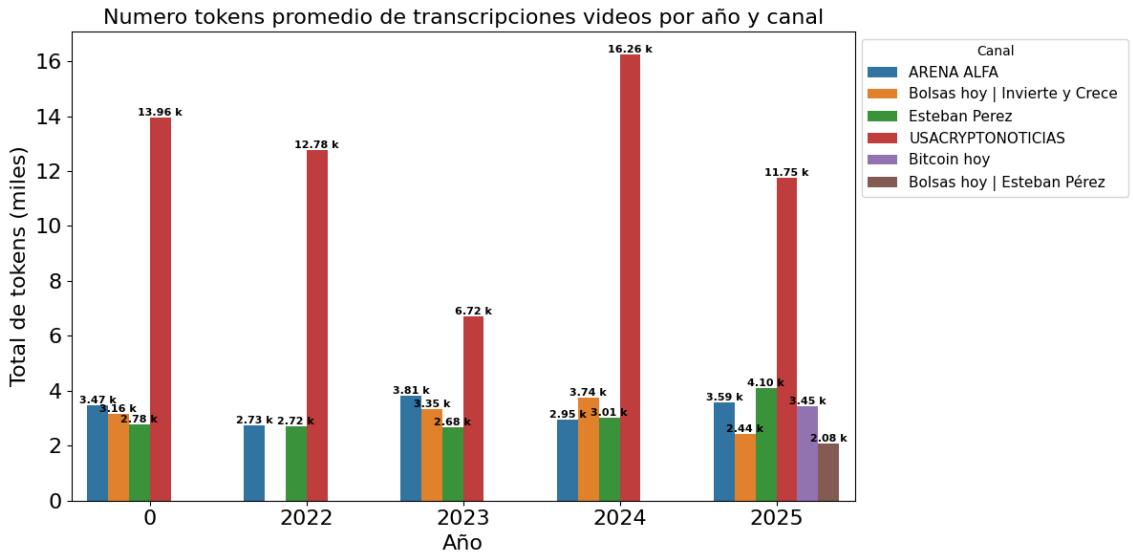

Figure 12: tokens de transcripciones videos publicados por año y canal.

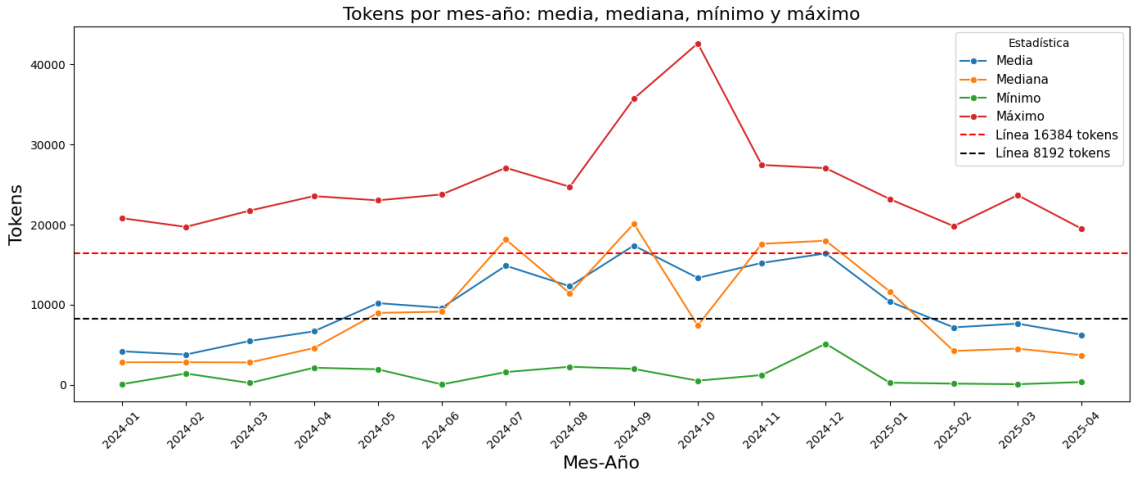

Figure 13: Tokens de transcripciones por mes y año para los años 2024 y 2025.

7.2 Experimentación

En la Tabla 3 se sintetizan los resultados obtenidos en la etapa de experimentación y ajuste fino de los modelos. A continuación, se describen los resultados en la experimentación más relevantes:

Uso de LoRA y OOM (Out of Memory): `LoRA` (Low-Rank Adaptation) es una técnica eficiente para adaptar grandes modelos de lenguaje. Permite trabajar con menos recursos, ya que sólo se ajusta una parte de los parámetros, utilizando de manera eficiente la RAM de la GPU mediante adaptadores de bajo rango y cuantización en 4 bits (`bnb.nf4`). Sin embargo, realizando el fine-tuning en modelos con mas de 8 billones de parámetros con recursos computacionales limitados, se produjo agotamiento de la memoria de la GPU (`OOM`, Out of Memory). Este problema se evidenció en la tabla comparativa: muchos experimentos, especialmente aquellos usando `Lightning-AI/litgpt` o configuraciones mayores a 4K de secuencia máxima, se quedaron sin memoria incluso antes de terminar el proceso de fine-tuning o durante la fase de validación.

El modelo **finetune_qlora_unsloth_llama_3_1_8B_ Instruct_bnb_4bit_v2** fue el mejor ajustado en el experimento ya que se logró completar el fine-tuning sin caer en `OOM` y alcanzó el menor error de entrenamiento (`train loss` = 1.655).

Limitaciones del proceso de validación: Durante el entrenamiento, no se calculó el error de validación ni la *perplexity* para los modelos más grandes, como el de 8B, principalmente porque estas métricas requieren procesar el conjunto de validación completo, lo cual demanda más memoria de GPU que la fase de entrenamiento normal, especialmente con secuencias largas.

Esto implica que no se realiza monitoreo del sobre-ajuste mediante validación en el ajuste fino del modelo por lo que no se calculan métricas de validación como `val_loss` o *perplexity*. Debido a que intentar calcular estas métricas se generan errores de `OOM`, interrumpiendo completamente la ejecución.

Ante esta limitación, la selección del mejor modelo se realizó utilizando el error de

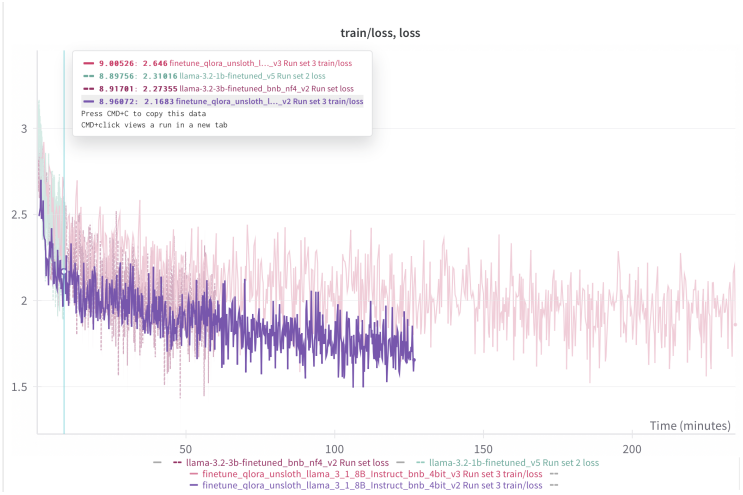

Figure 14: Error de entrenamiento mejores modelos ajustados usando LoRa.

Figure 15: Error entrenamiento mejor modelo ajustado usando LoRa.

entrenamiento (`train loss`) como referencia. Para más detalles, revisar el reporte publicado en el Anexo A.7.

Framework / Librería	**modelo base** (base model)	**Longitud máxima de secuencia** (max_seq_length)	**precisión cálculos** (precisión)	**cuantización modelo** (quantize)	**Parámetros ajustados**	**Memoria GPU GB** (Memory)	**Epocas** (Epochs)	**Tiempo (segundos)** (Time)	**batch global** (global_batch)	**Tamaño micro-batch** (micro_batch_size)	**Train loss**	**Error validación** (val loss)	**Perplexity validación** (val ppl)	**Nombre modelo entrenado** (Trained Model Name)
Unsloth	Llama-3.1-8B-Instruct	16384	bf16-true	bnb.nf4	41.943.040 (0,52%)	17,631	4	13.992	8	4	1,860	No se calcula	No se calcula	finetune_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v3
Unsloth	Llama-3.1-8B-Instruct	8192	bf16-true	bnb.nf4	41.943.040 (0,52%)	10,436	4	7.587	8	4	1,655	No se calcula	No se calcula	finetune_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v2
Lightning-AI/litgpt	Llama-3.2-3B-Instruct	> 4096	bf16-true	No	NA	OOM (Out of Memory)	NA	NA	1	1	NA	NA	NA	NA
Lightning-AI/litgpt	Llama-3.2-3B-Instruct	> 6000	bf16-true	bnb.nf4-dq	NA	OOM (Out of Memory)	NA	NA	1	1	NA	NA	NA	NA
Lightning-AI/litgpt	Llama-3.2-3B-Instruct	6000	bf16-true	bnb.nf4-dq	NA	OOM (Out of Memory)	4	NA	2	1	NA	NA	NA	NA
Lightning-AI/litgpt	Llama-3.2-3B-Instruct	6000	bf16-true	bnb.nf4-dq	2.293.760 (0,063%)	41,37	3	3.933	2	1	2,19	2,07	7,952	Llama-3.2-3B-Instruct-finetuned_bnb_nf4-dq_3
Lightning-AI/litgpt	Llama-3.2-3B-Instruct	6000	bf16-true	bnb.nf4	NA	OOM (Out of Memory)	4	NA	2	1	NA	NA	NA	NA
Lightning-AI/litgpt	Llama-3.2-3B-Instruct	6000	bf16-true	bnb.nf4-dq	2.293.760 (0,063%)	41,37	5	5.849,08	4	1	1,96866	2,07	7,935	Llama-3.2-3B-Instruct-finetuned_bnb_nf4-dq
Lightning-AI/litgpt	Llama-3.2-3B-Instruct	6000	bf16-true	bnb.nf4-dq	2.293.760 (0,063%)	41,37	5	5.583,41	8	1	1,76558	2,08	8,04	Llama-3.2-3B-Instruct-finetuned_bnb_nf4-dq_2
Lightning-AI/litgpt	Llama-3.2-3B-Instruct	4096	bf16-true	bnb.nf4	2.293.760 (0,063%)	29,18	5	3.525,29	16	1	2,017	2,11	8,26	llama-3.2-3b-finetuned_bnb_nf4_v3
Lightning-AI/litgpt	Llama-3.2-3B-Instruct	4096	bf16-true	bnb.nf4	2.293.760 (0,063%)	29,17	1	771,25	4	1	2,0647	2,19	8,95	llama-3.2-3b-finetuned_v1
Lightning-AI/litgpt	Llama-3.2-3B-Instruct	4096	bf16-true	bnb.nf4	2.293.760 (0,063%)	29,18	5	3.597,33	8	1	2,05171	2,08	8,03	llama-3.2-3b-finetuned_bnb_nf4_v2
Lightning-AI/litgpt	Llama-3.2-1B-Instruct	> 8192	bf16-true	No	NA	OOM (Out of Memory)	NA	NA	1	1	NA	NA	NA	NA
Lightning-AI/litgpt	Llama-3.2-1B-Instruct	4096	bf16-true	No	851.968 (0,057%)	36,17	1	350,18	2	1	2,30930	2.38	10,75	Llama-3.2-1B-finetuned_v2
Lightning-AI/litgpt	Llama-3.2-1B-Instruct	4096	bf16-true	No	851.968 (0,057%)	36,17	2	576,22	8	1	2,18275	2,38	10,82	Llama-3.2-1B-finetuned_v5
Lightning-AI/litgpt	Llama-3.2-1B-Instruct	4096	bf16-true	No	851.968 (0,057%)	36,17	1	438,74	1	1	2,24792	2,35	10,46	Llama-3.2-1B-finetuned

Table 3: Tabla comparativa experimentación entrenamiento modelos.

7.3 Resultados Cuantitativos evaluación

La evaluación de los modelos se realizó sobre un conjunto de prueba que comprendía 221 transcripciones de videos, representando el 10% del total de datos, balanceado por canal. La calidad de los resúmenes generados fue comparada con referencias producidas por el modelo fundacional GPT-4.1. Se utilizaron las siguientes métricas: distancia de coseno, un score simple (0-10) y un score basado en criterios específicos (fidelidad, relevancia, concisión y coherencia).

Los resultados comparativos de los modelos evaluados se sintetizan en la Tabla 5. La métrica inicial del modelo pre-entrenado Llama 3.1-8B-Instruct en configuración Zero-shot learning fue de 2.87 en el score promedio. En contraste, la métrica base, obtenida con el mismo modelo pero aplicando Few-shot learning (incluyendo un ejemplo en el prompt), mostró una mejora significativa, alcanzando un score promedio de 4.33. Este incremento representa una mejora del 50.6%. La aplicación del ajuste fino (Fine-tuning) mediante PEFT (LoRA) al modelo Llama 3.1-8B-Instruct (versión finetuned_qlora_unsloth_llama_3_1_8B _Instruct_bnb_4bit_v2_Q8_0), demostró un rendimiento superior, obteniendo un score promedio de 5.67. Esto constituye una mejora del 97.5% respecto a la métrica inicial del modelo sin ajuste fino. Este hallazgo subraya la eficacia del ajuste fino con LoRA en la mejora de la calidad de los resúmenes generados. Es notable que este modelo ajustado superó a modelos SLM con un mayor número de parámetros, como Phi4 de 14 billones de parámetros, que obtuvo un score de 4.51 en configuración Zero-shot learning.

El modelo finetune_qlora_unsloth_llama_3_1_8B _Instruct_bnb_4bit_v2_Q8_0 fue el seleccionado debido a su rendimiento óptimo y la finalización exitosa de su fine-tuning sin incurrir en errores de falta de memoria (OOM). Este modelo alcanzó el menor error de entrenamiento (train loss = 1.655). Se implementaron técnicas de optimización como la cuantización bnb.nf4 (4 bits) y precisión bf16 para optimizar el uso de memoria de la GPU.

Adicionalmente, se observó que la longitud máxima de contexto influye en las métricas de evaluación. Como se muestra en la Tabla 6, a mayor tamaño de contexto, las métricas promedio de criterial score, embedding cosine distance y score tienden a disminuir. Esto sugiere una degradación en el rendimiento del modelo a medida que aumenta la complejidad del texto de entrada.

7.4 Resultados Cualitativos

El proyecto tiene como objetivo generar reportes de análisis financiero similares a los elaborados por expertos, extrayendo y resumiendo información clave de transcripciones de videos. La generación de resúmenes se enfoca en el enfoque abstractivo, que permite condensar y reformular la información en un formato útil y comprensible para el usuario final. Un ejemplo ilustrativo de la estructura y contenido deseado para los reportes se presenta en la Tabla 4, donde se observa resumenes detallado generados mediante el modelo GPT-4.1 versus los generados con el modelo SLM ajustado. Este formato de reporte busca transformar la transcripción del video, que tiende a ser menos estructurado y a menudo incluye jerga técnica o interrupciones,

en un análisis corto coherente y accesible. Fácil de analizar en poco tiempo.

Resumen modelo Gpt 4.1	**Resumen modelo ajustado**	**Métricas**	**Numero tokens transcripción**
- **Introducción:** El texto proporciona un análisis técnico actualizado sobre la situación de Bitcoin al 5 de mayo, centrándose en la posible ruptura de su rango actual y evaluando tanto escenarios alcistas como bajistas. El análisis se basa en la acción del precio, patrones chartistas y niveles clave de soporte y resistencia, con especial atención a la importancia de la confirmación diaria y la reacción del mercado ante la apertura de Wall Street. - **Puntos clave:** - Se observa una figura de triples máximos en el gráfico diario de Bitcoin, lo que podría anticipar una corrección. - El cierre reciente se produjo ligeramente por debajo de los 28.963, nivel considerado relevante para confirmar un movimiento bajista. - Se requiere un segundo cierre diario en rojo para confirmar la probabilidad de una caída hacia niveles inferiores. - Durante la sesión asiática y la preapertura europea, el precio rebotó por encima de los 28.963, pero volvió a esa zona, manteniéndose en rango. - En el gráfico de 4 horas, se identifican dos escenarios: - **Escenario bajista:** Si se pierde el nivel de 28.963, el primer objetivo se sitúa en torno a los 25.900. - **Escenario alcista:** Si el precio rompe al alza, el objetivo relevante se encuentra en la zona de los 31.300, considerado un refugio de mediano plazo. - El análisis destaca la importancia de esperar la confirmación de la ruptura del rango antes de tomar decisiones de inversión. - Se anticipa un análisis más detallado que incluirá altcoins, índices y commodities en un próximo informe. - **Conclusión:** El análisis sugiere cautela ante la inminente ruptura del rango en Bitcoin, con niveles técnicos claramente definidos para ambos escenarios. La confirmación de la dirección del movimiento dependerá de la acción del precio en los próximos días, especialmente en torno al nivel de 28.963. Se recomienda a los inversores vigilar de cerca estos niveles y esperar confirmaciones antes de posicionarse. - **Activos recomendados:** - Bitcoin (BTC) - Altcoins (mencionadas para análisis futuro) - Índices (mencionados para análisis futuro) - Commodities (mencionados para análisis futuro)	- **Introducción:** El texto analiza la situación actual del mercado de Bitcoin, centrándose en la acción del precio, los niveles técnicos relevantes y posibles escenarios a corto plazo. Se observa un rango lateral con señales ambiguas, lo que exige cautela y vigilancia para determinar si el movimiento tendrá dirección bajista o alcista. - **Puntos clave:** - Bitcoin está en un rango lateral, formando una figura de triple máximo (posiblemente hombro-cabeza-hombro) en la serie diaria. - El cierre reciente fue por debajo de los 28.963 USD, pero no con suficiente margen para confirmar una corrección inmediata; requiere un segundo día consecutivo de caídas para reforzar el escenario bajista. - Durante la sesión asiática y preapertura europea, el precio mostró rebotes por encima de los 28.963 USD, pero posteriormente volvió a la zona inferior. - Se identifican dos posibles escenarios: - Bajista: Si se pierde el soporte de 28.963 USD, el siguiente objetivo bajista sería la zona de 25.900 USD (nivel referente en la serie de 4 horas). - Alcista: Un rompimiento al alza podría llevar el precio hacia el refugio de mediano plazo en los 31.300 USD. - El mercado se encuentra en una "trampa doble", lo que sugiere que ambos escenarios (bajista y alcista) podrían darse dependiendo de la apertura de Wall Street y el sentimiento del mercado. - **Conclusión:** El análisis destaca la alta incertidumbre actual en Bitcoin, con señales técnicas que pueden interpretarse tanto a favor como en contra de un movimiento bajista. La confirmación de uno u otro escenario depende principalmente de la reacción del precio ante la apertura de los mercados estadounidenses y la acción del precio en las próximas sesiones. Se recomienda cautela y seguimiento cercano de los niveles técnicos mencionados. - **Activos recomendados:** - Bitcoin (BTC)	score = 10.0 criterial_score = 1.0 cosine_distance = 0.096343	907

Resumen modelo Gpt 4.1	Resumen modelo ajustado	Métricas	Numero tokens transcripción
Informe de Análisis de Mercado — **Introducción** El texto analizado corresponde a un extenso y detallado análisis de mercado enfocado principalmente en Bitcoin (BTC) y el entorno de las criptomonedas, realizado por un analista con experiencia en trading. Se abordan proyecciones técnicas, estrategias operativas y reflexiones sobre la situación actual del mercado cripto, con énfasis en la importancia de la anticipación, la gestión de riesgos y la toma de decisiones fundamentadas en análisis técnico, especialmente en patrones como el hombro-cabeza-hombro y el comportamiento respecto a medias móviles relevantes. — **Puntos clave** - **Proyección bajista a corto/medio plazo para Bitcoin:** El analista sostiene que, tras un patrón de subida de 8 días, la probabilidad más alta es que BTC pierda la EMA 200 en gráfico diario, con caídas proyectadas hacia la zona de 74,000-68,000 USD, e incluso menciona la posibilidad de llegar a 67,000 USD para liquidar posiciones apalancadas. - **Patrón técnico dominante:** Se destaca la formación de un hombro-cabeza-hombro (HCH) en el gráfico semanal de BTC, lo que sugiere un posible cambio de ciclo de alcista a bajista si se confirma un “alto menor” en el rebote proyectado. - **Estrategia operativa recomendada:** El enfoque es operar con base en anticipación y no en reacción. El analista recomienda estar preparado para ambos escenarios (caída y rebote), acumulando posiciones en spot en zonas de soporte y gestionando shorts abiertos desde zonas altas. - **Gestión de posiciones y riesgo:** Se enfatiza la importancia de tomar ganancias parciales, promediar compras en caídas y evitar operar con el “corazón”. Se recomienda operar en spot para inversores de largo plazo y considerar la toma de beneficios en shorts ante rebotes técnicos. - **Proyección de rebote tras la caída:** Tras la caída esperada, se proyecta un rebote en el gráfico semanal, que podría llevar a BTC a la zona de 95,000-100,000 USD. Sin embargo, si este rebote resulta en un “alto menor” que el anterior, se confirmaría un cambio de estructura bajista, con posibles caídas hacia 40,000-30,000 USD. - **Importancia de la EMA 200 y la “mata cicos”:** La pérdida de la EMA 200 en semanal es vista como señal de ciclo bajista. La “mata cicos” (presumiblemente una resistencia clave) es mencionada como nivel crítico que, de no superarse, confirmaría la debilidad del mercado. - **Situación de las altcoins:** Se advierte que las altcoins han sufrido caídas significativas (40-70% en muchos casos) y que no se espera una “altseason” independiente; su recuperación dependerá del rebote de BTC. - **Recomendaciones de acumulación:** Se sugiere que los mejores precios de compra para largo plazo suelen encontrarse cuando los activos están por debajo de la EMA 200 semanal, tanto para BTC como para altcoins relevantes.	**Resumen de Conversación** - El tema principal de la conversación es la percepción de millonarios y sus hábitos en relación con otros influyentes del mercado financiero, especialmente en el ámbito de trading e inversiones. - Se destaca que personas muy afortunadas no suelen invertir su tiempo observando o criticar a otros, ya que prefieren centrarse en actividades productivas y útiles. - Se menciona la diversión que supone identificar a quienes se dedican a difundir mentiras sobre sus ganancias o perfiles financieros, sugiriendo una actitud desconfiada hacia los ”héroes de trading” poco serios. **Puntos clave para inversores:** 1. **No te distraigas**: No pierdas tiempo leyendo historias fantasosas de otros, especialmente aquellas que parecen imposibles o exageradas. 2. **Focus en tus inversiones**: Mantén tu atención en actividades útiles y productivas, evitando la distracción por rumores o mentiras de terceros. 3. **Cuidado con las redes sociales**: En el entorno digital, es fácil que aparezcan historias falsas o exageradas sobre ganancias; no te crees todo lo que se publica sin evidencia. **Contexto adicional:** - La conversación se centra en la psicología y comportamiento de quienes acumulan grandes fortunas financieras. - Se mencionan redes sociales y plataformas digitales como lugares donde se difunden historias atractivas pero posiblemente falsas sobre ganancias o perfiles financieros. **Recomendaciones para inversores:** - Verifica la información antes de tomar decisiones basadas en rumores o anécdotas. - Mantén una actitud crítica y desconfiada frente a historias aparentemente demasiado buenas para ser verdad. - Centra tu atención en estrategias comprobadas y análisis fundamentales. Este texto resalta la importancia de no dejarse llevar por rumores o anécdotas falsas, recomendando un enfoque prudente y basado en evidencia en las decisiones de inversión.	score = 1.0 criterial_score = 0.0 cosine_distance = 0.421213	25516

Resumen modelo Gpt 4.1	Resumen modelo ajustado	Métricas	Numero tokens transcripción
- **Contexto macro y factores externos:** Se desestima la influencia de factores externos como declaraciones de Donald Trump, aranceles o recesión, priorizando el análisis técnico sobre la narrativa de noticias. - **Gestión emocional y aprendizaje:** Se aconseja operar sin emociones, aprender de los ciclos y prepararse para el próximo, destacando la importancia de la experiencia y la formación continua. — **Conclusión** El análisis expone una visión técnica y estratégica del mercado cripto, con énfasis en la anticipación y la gestión de riesgos. Se proyecta una caída adicional en BTC y altcoins, seguida de un rebote técnico que podría definir el cambio de ciclo. La recomendación principal es operar con disciplina, aprovechar las zonas de compra en spot para largo plazo y gestionar activamente las posiciones en futuros y shorts. Se desaconseja operar por FOMO o influencias externas, y se invita a los inversores a formarse y operar con base en análisis técnico y gestión emocional. — **Activos recomendados** Según el texto, los activos mencionados como opciones de inversión (para spot, acumulación o trading) son: - **Bitcoin (BTC)** - **Ethereum (ETH)** - **Solana (SOL)** - **TRON (TRX)** - **BNB (Binance Coin)** - **XRP** - **CAKE (PancakeSwap)** - **Litecoin (LTC)** - **LINK (Chainlink)** - **ADA (Cardano)** - **Hedera Hashgraph (HBAR)** - **RUNE** - **STX** - **OP** - **PEPE** - **ONDO** - **IOTA** - **SAND** - **JUPITER** - **TIA** - **CORE** - **INJ** - **MOVR** - **STPT** **Nota:** La mención de estos activos no implica una recomendación de compra inmediata, sino que han sido referenciados en el análisis como opciones para seguimiento, acumulación en zonas de soporte o ejemplos de estrategias de trading. Se recomienda realizar análisis adicional antes de invertir.			

Resumen modelo Gpt 4.1	Resumen modelo ajustado	Métricas	Numero tokens transcripción
- **Introducción:** El texto analiza de manera crítica la pregunta frecuente sobre cuánto dinero se necesita para vivir del trading o de las inversiones en bolsa, desmitificando las falsas promesas de libertad financiera y rentabilidades elevadas que suelen circular en redes sociales. Se enfatiza la importancia de un enfoque realista, basado en la experiencia, la rentabilidad histórica y la gestión del riesgo, para quienes consideran dedicarse profesionalmente al trading. - **Puntos clave:** - Existen muchas falsas promesas en redes sociales sobre la facilidad de vivir del trading, asociadas a imágenes de éxito superficial y esquemas fraudulentos como pirámides o multiniveles. - No hay una cifra fija para vivir del trading; depende de los gastos personales, la rentabilidad histórica demostrada y la capacidad de soportar meses o años malos. - La rentabilidad en los mercados no es fija ni garantizada. Los mejores inversionistas del mundo, como Warren Buffett, George Soros y Peter Lynch, han logrado rentabilidades promedio anuales entre el 19% y el 39%, lo que equivale aproximadamente a un 1% a 3% mensual. - Es irreal esperar vivir del trading con pequeños capitales (por ejemplo, $10,000 o $50,000 pesos, o $1,000 a $5,000 dólares). Se requieren capitales significativamente mayores (por ejemplo, $500 millones de pesos colombianos, equivalentes a unos $125,000 dólares) para obtener ingresos mensuales viables, incluso asumiendo rentabilidades históricas de los mejores gestores. - Las rentabilidades fijas mensuales ofrecidas en productos de renta variable, forex o criptomonedas son una señal clara de fraude. - El trading es una carrera compleja que requiere años de estudio, disciplina y experiencia. Solo una minoría logra el éxito sostenido. - Es fundamental tener un plan para afrontar periodos de pérdidas, especialmente si se depende exclusivamente del trading para vivir. - **Conclusión:** Vivir del trading es un objetivo alcanzable solo para una minoría disciplinada, con experiencia comprobada y un capital considerable. Las promesas de libertad financiera rápida y rentabilidades fijas elevadas son engañosas y, en muchos casos, fraudulentas. Los inversores deben basar sus expectativas en datos históricos reales y en un análisis honesto de su situación financiera y capacidad de gestión del riesgo. - **Activos recomendados:** El texto menciona los siguientes activos y mercados como opciones de inversión (sin recomendarlos explícitamente, sino como ejemplos de lo que se suele ofrecer o analizar): - Acciones (bolsa de valores) - Forex (mercado de divisas) - Bitcoin (criptomonedas) - Fondos de inversión **Nota:** El texto no recomienda ningún activo específico, sino que los menciona en el contexto del análisis y advertencia sobre falsas promesas de rentabilidad.	- **Introducción:** El texto analiza de manera crítica la pregunta frecuente sobre cuánto dinero se necesita para vivir del trading o de las inversiones en bolsa, desmitificando las falsas promesas de libertad financiera y rentabilidades elevadas que suelen circular en redes sociales. Se enfatiza la importancia de un enfoque realista, basado en la experiencia, la rentabilidad histórica y la gestión del riesgo, para quienes consideran dedicarse profesionalmente al trading. - **Puntos clave:** - Existen muchas falsas promesas en redes sociales sobre la facilidad de vivir del trading, asociadas a imágenes de éxito superficial y esquemas fraudulentos como pirámides o multiniveles. - No hay una cifra fija para vivir del trading; depende de los gastos personales, la rentabilidad histórica demostrada y la capacidad de soportar meses o años malos. - La rentabilidad en los mercados no es fija ni garantizada. Los mejores inversionistas del mundo, como Warren Buffett, George Soros y Peter Lynch, han logrado rentabilidades promedio anuales entre el 19% y el 39%, lo que equivale aproximadamente a un 1% a 3% mensual. - Es irreal esperar vivir del trading con pequeños capitales (por ejemplo, $10,000 o $50,000 pesos, o $1,000 a $5,000 dólares). Se requieren capitales significativamente mayores (por ejemplo, $500 millones de pesos colombianos, equivalentes a unos $125,000 dólares) para obtener ingresos mensuales viables, incluso asumiendo rentabilidades históricas de los mejores gestores. - Las rentabilidades fijas mensuales ofrecidas en productos de renta variable, forex o criptomonedas son una señal clara de fraude. - El trading es una carrera compleja que requiere años de estudio, disciplina y experiencia. Solo una minoría logra el éxito sostenido. - Es fundamental tener un plan para afrontar periodos de pérdidas, especialmente si se depende exclusivamente del trading para vivir. - **Conclusión:** Vivir del trading es un objetivo alcanzable solo para una minoría disciplinada, con experiencia comprobada y un capital considerable. Las promesas de libertad financiera rápida y rentabilidades fijas elevadas son engañosas y, en muchos casos, fraudulentas. Los inversores deben basar sus expectativas en datos históricos reales y en un análisis honesto de su situación financiera y capacidad de gestión del riesgo.	score = 6.0 criterial_score = 1.0 cosine_distance = 0.172059	3852

Resumen modelo Gpt 4.1	Resumen modelo ajustado	Métricas	Numero tokens transcripción
	- **Activos recomendados:** El texto menciona los siguientes activos y mercados como opciones de inversión (sin recomendarlos explícitamente, sino como ejemplos de lo que se suele ofrecer o analizar): - Acciones (bolsa de valores) - Forex (mercado de divisas) - Bitcoin (criptomonedas) - Fondos de inversión **Nota:** El texto no recomienda ningún activo específico, sino que los menciona en el contexto del análisis y advertencia sobre falsas promesas de rentabilidad.		

Table 4: Resumen referencia y resumen generado por modelo ajustado.

Comparación de resúmenes: En el Caso 1, que corresponde a una transcripción de 907 tokens, el modelo ajustado obtuvo métricas destacadas: un score de 10.0, un score criterial de 1.0 y una distancia de coseno de 0.096343. Estos valores indican una alta similitud y calidad en comparación con el resumen de referencia generado por GPT-4.1.

Fortalezas del Modelo en Contextos Moderados (transcripciones cortas, es decir videos no muy extensos): Fidelidad a la información clave: El modelo ajustado demuestra una notable capacidad para identificar y reproducir los puntos neurálgicos del análisis técnico de Bitcoin, incluyendo los niveles críticos de soporte y resistencia (28.963 USD, 25.900 USD, 31.300 USD) y los patrones chartistas (figura de triple máximo o hombro-cabeza-hombro).

Concisión y Coherencia: A pesar de las características no estructuradas del lenguaje hablado en las transcripciones, el modelo genera un resumen que mantiene una estructura lógica y una fluidez que facilita su lectura rápida y comprensible.

Uso adecuado de tecnicismos: El modelo integra eficazmente la jerga técnica financiera, como "rango lateral," "soportes," "rompimiento al alza," y "hombro-cabeza-hombro," lo cual es crucial para la precisión y utilidad de un reporte de análisis financiero destinado a inversores. Esto reduce la necesidad de post-edición para asegurar la terminología correcta.

Análisis de Escenarios: La capacidad de delinear claramente los escenarios bajista y alcista con sus respectivos objetivos de precio es una fortaleza clave, proporcionando al inversor una visión estructurada de las posibles trayectorias del mercado.

En contraste, el Caso 2, que involucra una transcripción considerablemente más extensa de 25516 tokens, revela una debilidad crítica en la capacidad del modelo para mantener la coherencia temática. Mientras que el resumen generado por el modelo GPT-4.1 se centra en un "Informe de Análisis de Mercado" detallado sobre Bitcoin y criptomonedas, el resumen del modelo ajustado aborda la "percepción de millonarios y sus hábitos en relación con otros influyentes del mercado financiero".

Debilidades del Modelo en Contextos Extendidos (transcripciones largas, es decir videos muy extensos): Error de Interpretación Temática o "Deriva Temática": La disparidad temática entre el resumen de referencia (análisis técnico de mercado) y el generado por el modelo ajustado (hábitos de millonarios) constituye una debilitación sustancial del modelo en el manejo de contextos extendidos. A pesar de que las métricas asociadas al resumen generado por el modelo ajustado (score $\bar{1}0.0$, criterial_score $\bar{1}.0$, cosine_distance $\bar{0}.421213$) indican una alta calidad intrínseca del texto generado en relación con su propio contenido, no reflejan la fidelidad al tema principal de la transcripción original tal como fue interpretado por el modelo de referencia GPT-4.1. Este resultado sugiere una incapacidad del modelo ajustado para mantener la coherencia temática o identificar el mensaje central en transcripciones de mayor longitud, lo que podría conducir a la generación de resúmenes irrelevantes para el propósito financiero buscado.

Degradación del Rendimiento con la Longitud del Contexto: Este hallazgo es consistente con los resultados cuantitativos generales que indican que a mayor tamaño de contexto (número de tokens), las métricas de evaluación promedio (criterial score, embedding cosine distance y score) tienden a disminuir. En este caso particular, la degradación se manifiesta no solo como una disminución en las métricas de calidad general, sino como un cambio fundamental en el tema, lo cual es más crítico que una simple reducción en la precisión o concisión, comprometiendo la utilidad del reporte financiero.

En resumen, el modelo ajustado demuestra una gran capacidad para generar resúmenes cualitativos de alta calidad y fidelidad temática a partir de transcripciones de videos de análisis financiero de longitud moderada, logrando una comprensión y síntesis abstractiva efectiva. No obstante, el análisis de casos con transcripciones considerablemente más extensas revela la necesidad de optimizar el rendimiento en contextos de mayor longitud. Es crucial refinar la capacidad del modelo para capturar la totalidad de los matices y advertencias críticas emitidas por los expertos y, fundamentalmente, mantener la coherencia temática principal del contenido original, aspectos esenciales para la toma de decisiones informadas en un mercado volátil.

7.5 Despliegue

A continuación se describe las funcionalidades desplegados de la herramienta:

Reporte de video: Permite al usuario ver el listado de videos de canales en los últimos 7 días, adicionalmente permite al usuario seleccionar la url de uno de ellos o pegar una url de un video cualquiera, seleccionar el modelo y el prompt para generar el reporte resumido. Ver Figura 16 y Figura 17.

Reporte de métricas: Permite seleccionar las métricas definidas para comparar los distintos modelos en el conjunto de evaluación y ver su performance actual. Ver Figura 18.

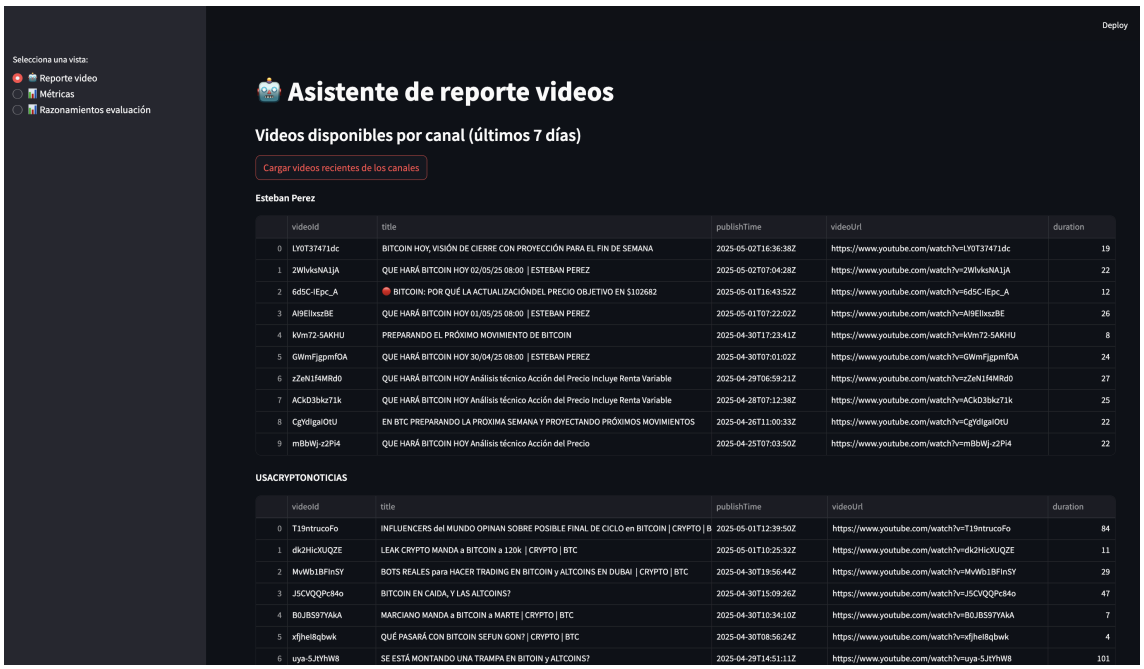

Figure 16: Sección de videos interfaz gráfica.

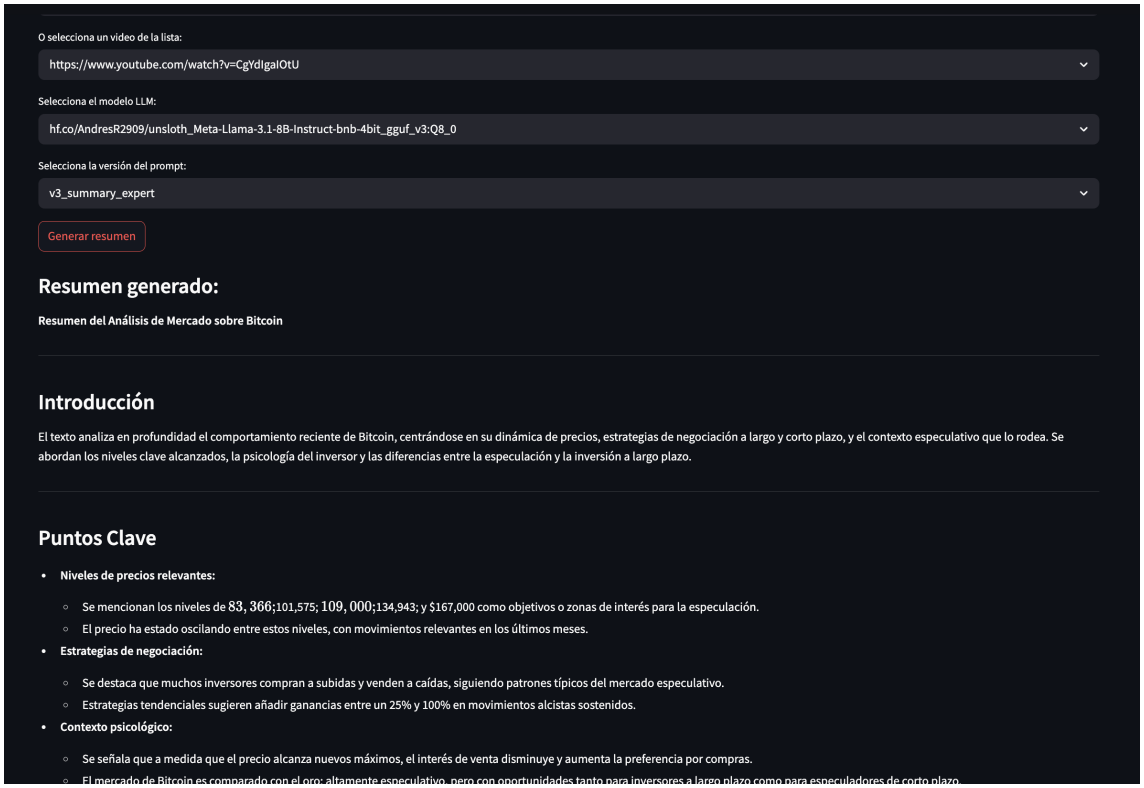

Figure 17: Sección de reporte de video interfaz gráfica.

Figure 18: Sección de reporte métricas interfaz gráfica.

7.6 Reflexión final

Los resultados obtenidos demuestran que el proyecto ha avanzado significativamente en el cumplimiento de su objetivo general: desarrollar una herramienta que asista a los inversores individuales mediante la generación de reportes resumidos personalizados de análisis de expertos en tiempo oportuno. La utilización de SLMs ajustados mediante LoRA ha probado ser una estrategia costo-eficiente para democratizar el acceso a información financiera experta.

Fortalezas y logros clave:

Democratización del acceso a información: La herramienta permite a inversores individuales acceder a análisis de expertos de manera oportuna, reduciendo la brecha con inversores institucionales.

Optimización de costos y recursos: El uso de SLMs y la técnica LoRA optimiza los costos de implementación del sistema en comparación con LLMs fundacionales de gran escala. Por ejemplo, se estima un ahorro anual de 7.52 a 25 dólares en costos de API para los canales procesados en esta prueba de concepto, con un potencial de mayor relevancia a gran escala. Ver Figura 19 y Figura 20

Ahorro de tiempo significativo: La herramienta puede ahorrar a los inversores individuales un promedio de 560 horas al año (497 horas en 2024 y 204 horas en el primer semestre de 2025 para los canales seleccionados) en el consumo de análisis de expertos.

Eficiencia del ajuste fino con LoRA: LoRA ha demostrado ser altamente recomendable para adaptar LLMs a tareas específicas sin incurrir en los elevados costos computacionales del full fine-tuning, haciendo el proceso más rápido, barato

y escalable.

Limitaciones identificadas:

Rendimiento en comparación con LLMs propietarios: Aunque el modelo ajustado muestra mejoras notables, no logró igualar el rendimiento del modelo propietario GPT-4o-mini. Para videos cortos, el modelo ajustado se sitúa 2.4 puntos por debajo en la métrica de score.

Restricciones de memoria (OOM) y validación durante el entrenamiento: Durante la fase de experimentación, se enfrentaron problemas de falta de memoria (OOM) en configuraciones con secuencias máximas mayores a 4K, especialmente al intentar calcular métricas de validación en modelos más grandes. Esto impidió un monitoreo en tiempo real del sobreajuste mediante val_loss o perplexity.

Impacto de la longitud del contexto: Se observó que a mayor longitud de contexto de las transcripciones, las métricas de evaluación de los modelos tendían a disminuir. Esto sugiere que la complejidad de las entradas largas sigue siendo un desafío.

Posibles líneas de mejora para trabajos futuros:

Estrategias para manejo de contexto largo: Para transcripciones que excedan los 8192 tokens, se podría explorar la técnica de "map reduce" o "refinamiento interactivo". Esto implicaría dividir las transcripciones en "chunks", resumir cada uno, y luego generar un resumen consolidado, lo que podría mejorar el rendimiento del modelo en contextos más extensos.

Modelo híbrido de despliegue: Una estrategia de implementación podría involucrar un modelo mixto, utilizando la API de GPT-4o-mini para videos extensos (comunes en el canal Criptonoticias) y el modelo ajustado para videos más cortos (24 minutos o menos), optimizando así los costos de operación.

Expansión del conjunto de datos: La adición de más datos de entrenamiento, potencialmente generados con técnicas avanzadas, podría mejorar la capacidad de generalización y el rendimiento de los modelos.

Optimización de recursos computacionales: Continuar explorando configuraciones de hiperparámetros y técnicas de cuantificación más eficientes para soportar contextos de mayor longitud y modelos de mayor tamaño sin incurrir en OOM.

En resumen, los resultados obtenidos evidencian que el ajuste fino de modelos SLM mediante técnicas como LoRA permite generar resúmenes financieros con alta similitud semántica respecto a referencias generadas por modelos fundacionales como GPT-4.1, y con una evaluación cualitativa favorable por parte de LLM-as-a-judge. El modelo seleccionado ofrece un equilibrio adecuado entre rendimiento y eficiencia computacional, facilitando su despliegue en una aplicación web funcional. Estos resultados confirman la viabilidad técnica y profesional de la solución propuesta, alineándose con los objetivos del proyecto y destacando su potencial de escalabilidad en escenarios reales de análisis financiero.

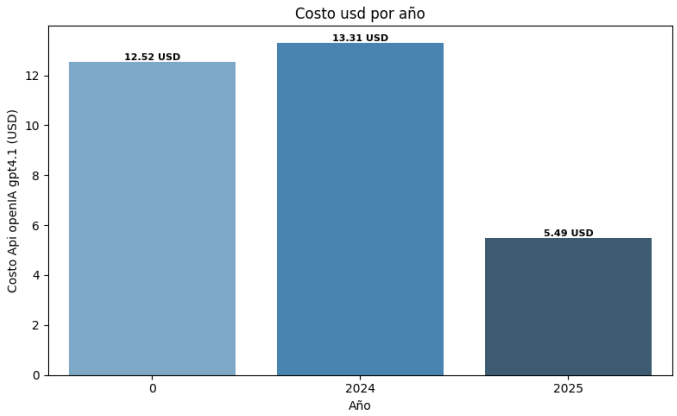

Figure 19: Costos anuales aproximados generación de reportes resumidos usando gpt4.1.

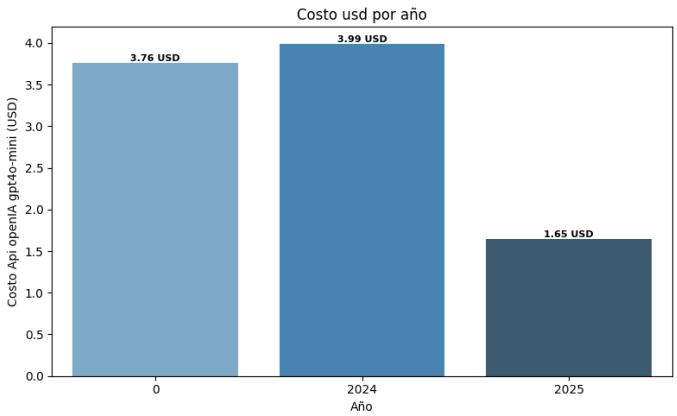

Figure 20: Costos anuales aproximados generación de reportes resumidos usando gpt4o-mini.

Modelo	Learning	Promedio criterio score	Promedio embedding cosine distance	Promedio score	Observación
gpt_4o_mini	Zero shot learning	0,602	0,101	8,48	Métrica modelo 4o-mini Open IA (API)
finetuned_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v3_Q8_0	Fine-tuning	0,294	0,159	5,98	
finetuned_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v2_Q8_0	**Fine-tuning**	**0,367**	**0,159**	**5,67**	**Modelo seleccionado**
finetuned_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v3_Q4_k_m	Fine-tuning	0,262	0,174	5,19	
phi4_latest	Few shot learning	0,167	0,186	5,18	
finetuned_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v2_Q4_k_m	Fine-tuning	0,285	0,179	5,15	
phi4_latest	**Zero shot learning**	**0,172**	**0,197**	**4,51**	**Métrica de referencia:** modelo 14 Billones de parámetros cuantizado a 4 bits.
llama3_1_8B_instruct_fp16	**Few shot learning**	**0,081**	**0,203**	**4,33**	**Métrica base**
finetune_qlora_litgpt_llama_3_2_3B_bnb_nf4_v2_q8_0	Fine-tuning	0,140	0,225	3,33	
llama3_2_3B_instruct_fp16	Few shot learning	0,027	0,226	3,21	
llama3_1_8B_instruct_fp16	**Zero shot learning**	**0,050**	**0,383**	**2,87**	**Métrica inicial:** Modelo de referencia modelo 8 Billones de parámetros cuantizado a 16 bits.
llama3_2_3B_instruct_fp16	Zero shot learning	0,027	0,428	2,10	

Table 5: Resultados de evaluación de modelos con métricas definidas

Modelo	Learning	Promedio criterial score	Promedio embedding cosine distance	Promedio score
Longitud de contexto máxima de 4096				
gpt_4o_mini	Zero shot learning	0.685	0.087	8.73
finetune_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v2_Q8_0	Fine-tuning	0.516	0.135	6.81
finetune_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v2_Q4_k_m	Fine-tuning	0.427	0.151	6.31
llama3_1_8b_instruct_fp16	Few shot learning	0.145	0.180	5.13
llama3_1_8b_instruct_fp16	Zero shot learning	0.081	0.310	3.65
Longitud de contexto máxima de 8192				
gpt_4o_mini	Zero shot learning	0.646	0.090	8.68
finetune_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v2_Q8_0	Fine-tuning	0.468	0.145	6.40
finetune_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v2_Q4_k_m	Fine-tuning	0.367	0.164	5.86
llama3_1_8b_instruct_fp16	Few shot learning	0.114	0.190	4.82
llama3_1_8b_instruct_fp16	Zero shot learning	0.070	0.314	3.37
Longitud de contexto máxima de 16384				
gpt_4o_mini	Zero shot learning	0.625	0.094	8.65
finetune_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v2_Q8_0	Fine-tuning	0.432	0.150	6.23
finetune_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v2_Q4_k_m	Fine-tuning	0.335	0.169	5.66
llama3_1_8b_instruct_fp16	Few shot learning	0.102	0.194	4.65
llama3_1_8b_instruct_fp16	Zero shot learning	0.063	0.336	3.20

Table 6: Resultados de evaluación de modelos con diferentes longitudes máximas de contexto

8 Conclusiones

El presente proyecto de investigación ha culminado exitosamente en el desarrollo e implementación de una herramienta computacional avanzada, específicamente diseñada para optimizar la toma de decisiones financieras de inversores individuales. Esta solución innovadora se sustenta en el ajuste fino de un *Small Language Model* (SLM) y la integración sistemática de transcripciones de contenido audiovisual de expertos disponible en plataformas como YouTube, abordando de manera directa los desafíos inherentes a la sobrecarga informacional y la complejidad analítica de datos no estructurados en el dinámico mercado financiero contemporáneo.

El análisis subsiguiente de la ejecución del proyecto evidencia que los objetivos preestablecidos han sido satisfactoriamente alcanzados, lo que marca un avance significativo hacia la democratización del acceso a información financiera especializada y la mejora de la capacidad analítica de los usuarios:

- Se logró la integración y procesamiento de datos provenientes de fuentes no estructuradas, específicamente transcripciones de videos de análisis de expertos de YouTube. Esto permite la extracción de información relevante y actualizada, fundamental para la toma de decisiones en un entorno dinámico.

- Se ha logrado el ajuste fino exitoso de un modelo SLM, específicamente el modelo `finetuned_qlora_unsloth_llama_3_1_8B_Instruct_bnb_4bit_v2_Q8_0`. Este modelo ajustado alcanzó un puntaje promedio de $5,67$ en la evaluación automática realizada por un LLM-as-a-judge (GPT-4.1) y un *criterial score* promedio de $0,367$, con una distancia coseno promedio de $0,159$. Estos resultados representan una mejora del $97,5\,\%$ respecto a la métrica inicial del modelo base sin ajuste fino, demostrando su capacidad para generar resúmenes coherentes y útiles para inversores individuales. Además, superó en rendimiento a modelos de mayor escala como `Phi4` (14B parámetros) en configuración *zero-shot*.

- La evaluación rigurosa de la calidad de los reportes generados se fundamentó en métricas cuantitativas objetivas, como la similitud de coseno entre *embeddings* y las puntuaciones generadas por Modelos de Lenguaje Grandes (LLMs) operando como evaluadores (*LLM-as-a-judge*). Esta metodología es respaldada por la alta correlación documentada entre las valoraciones de los LLMs y las evaluaciones humanas en tareas de generación textual.

- El despliegue del modelo sobre una aplicación web asegura la accesibilidad y facilidad de uso para el usuario final, permitiendo la generación de reportes resumidos a partir de videos de análisis de expertos en *trading*. Este componente es crucial para que la herramienta pueda proporcionar *insights* personalizados de manera oportuna.

- Desde una perspectiva de eficiencia operativa, la implementación de esta herramienta se traduce en un ahorro cuantificable de tiempo y costos para los

inversores individuales. Se estima un ahorro promedio de aproximadamente 560 horas anuales en la revisión de análisis de expertos para los canales seleccionados en este estudio, junto con una reducción de costos de API de entre $7,52 y $25 anuales para la misma muestra, lo cual, a mayor escala de canales y videos procesados, se traduciría en una relevancia económica significativa.

No obstante los logros alcanzados, la implementación de este sistema ha puesto de manifiesto ciertas limitaciones inherentes al dominio y a las tecnologías empleadas:

- La complejidad intrínseca de la extracción y síntesis de información de fuentes no estructuradas, como las transcripciones de videos, persiste como un desafío principal, dado que el lenguaje hablado es inherentemente menos estructurado y puede contener repeticiones, jerga específica y contexto visual no textualizado.
- La calidad y diversidad del conjunto de datos de entrenamiento constituyeron una limitación, afectando potencialmente la capacidad de generalización del modelo, lo que subraya la criticidad de corpus especializados para el óptimo desarrollo de modelos de lenguaje en dominios específicos.
- A pesar de la mejora en la cuantificación de la calidad mediante *LLMs-as-a-judge* y similitud de coseno, la evaluación de resúmenes abstractivos mantiene una componente de subjetividad, lo que presenta un reto continuo.
- Se observó una degradación en el rendimiento del modelo ajustado al procesar transcripciones de mayor longitud. Aunque el modelo ajustado superó al `Phi4`, no igualó el rendimiento del modelo propietario *GPT-4o-mini*, situándose $2,4$ puntos por debajo en la métrica de *score* para videos cortos.

Para potenciar la capacidad y maximizar el impacto de este sistema, se han delineado las siguientes líneas estratégicas para futuras investigaciones y desarrollos:

- **Expansión y diversificación de conjuntos de datos:** La ampliación del corpus de entrenamiento y la exploración de dominios temáticos más heterogéneos dentro del contenido financiero podrían mejorar sustancialmente la robustez del modelo.
- **Exploración de arquitecturas y técnicas avanzadas:** Aunque LoRA demostró eficiencia, se propone investigar la integración de otras técnicas de *PEFT* o marcos como el *Retrieval Augmented Generation* (RAG) para un análisis más contextualizado y una mitigación de la "alucinación". Adicionalmente, la experimentación con el aprendizaje por refuerzo con retroalimentación humana (*RLHF*) podría perfeccionar el rendimiento del modelo.
- **Integración de validación humana exhaustiva:** Si bien los métodos automáticos son eficientes, la incorporación de un proceso de validación humana es indispensable para asegurar la utilidad práctica y la comprensibilidad de los resúmenes para la audiencia no experta.

- **Optimización para contextos extensos y modelos híbridos:** Dada la observación de un rendimiento inferior en transcripciones de mayor longitud, se plantea la implementación de estrategias como 'map reduce' o 'refinamiento interactivo' para procesar *chunks* de texto extensos y generar resúmenes consolidados (ver Anexo A.6). Asimismo, se considera un modelo de despliegue híbrido, utilizando la API de *GPT-4o-mini* para videos muy extensos (e.g., del canal `CRIPTONOTICIAS`) y el modelo SLM ajustado para videos de menor duración ($\leq$ 24 minutos), lo que podría equilibrar el rendimiento y la eficiencia de costos.

Como conclusión final, la herramienta desarrollada exhibe un gran potencial transformador en la interacción de los inversores individuales con la información financiera. Al generar reportes eficientes, accesibles y concisos que destilan información clave en lenguaje natural, el sistema asiste directamente en la toma de decisiones fundamentadas a partir de datos no estructurados. Su diseño para ser desplegado como una aplicación web subraya su orientación práctica y su capacidad de integración en entornos operativos reales, facilitando el acceso oportuno a análisis expertos y contribuyendo a la democratización de la información financiera especializada. La optimización de costos mediante SLMs robustece su viabilidad para despliegues con recursos limitados, posicionándola como una solución efectiva y práctica para el inversor individual, acercándose al objetivo de proporcionar herramientas que mejoren la capacidad de respuesta y decisión en un mercado en constante evolución.

9 Referencias

References

Bandi, A., & Kagitha, H. (2024). A Case Study on the Generative AI Project Life Cycle Using Large Language Models, 189–177. https://doi.org/10.29007/hvzc

Chen, D. (2023). Deep Learning-Based Investor Sentiment Indicator Construction and Stock Prediction Research. *2023 3rd International Signal Processing, Communications and Engineering Management Conference (ISPCEM)*, 31–35. https://doi.org/10.1109/ISPCEM60569.2023.00012

Chen, Y.-P., Chu, K., & Nakayama, H. (2024, June). LLM as a Scorer: The Impact of Output Order on Dialogue Evaluation [arXiv:2406.02863 [cs]]. https://doi.org/10.48550/arXiv.2406.02863

Dong, M. M., Stratopoulos, T. C., & Wang, V. X. (2024). A scoping review of ChatGPT research in accounting and finance. *International Journal of Accounting Information Systems*, *55*, 100715. https://doi.org/10.1016/j.accinf.2024.100715

Du, K., Zhao, Y., Mao, R., Xing, F., & Cambria, E. (2025). Natural language processing in finance: A survey. *Information Fusion*, *115*, 102755. https://doi.org/10.1016/j.inffus.2024.102755

Gao, D., Ma, Y., Liu, S., Song, M., Jin, L., Jiang, W., Wang, X., Ning, W., Yu, S., Xuan, Q., Cai, X., & Yang, L. (2024). FashionGPT: LLM instruction fine-tuning with multiple LoRA-adapter fusion. *Knowledge-Based Systems*, *299*, 112043. https://doi.org/10.1016/j.knosys.2024.112043

Grauer, G. (2024). Infront Analytics. https://www.infrontanalytics.com/

Gu, Y., Dong, L., Wei, F., & Huang, M. (2024, April). MiniLLM: Knowledge Distillation of Large Language Models [arXiv:2306.08543 [cs]]. https://doi.org/10.48550/arXiv.2306.08543

Han, J., Kamber, M., & Pei, J. (2012). Getting to Know Your Data. In *Data Mining* (pp. 39–82). Elsevier. https://doi.org/10.1016/B978-0-12-381479-1.00002-2

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021, October). LoRA: Low-Rank Adaptation of Large Language Models [arXiv:2106.09685 [cs]]. Retrieved November 12, 2024, from http://arxiv.org/abs/2106.09685

Huang, B., Wang, L., & Zhou, L. (2022). Information Retrieval from Internet Data Using Textual Mining and Statistical Models for Stock Market Price Movement Prediction. *2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*, 308–312. https://doi.org/10.1109/TOCS56154.2022.10015918

Ijebu, F. F., Liu, Y., Sun, C., & Usip, P. U. (2025). Soft cosine and extended cosine adaptation for pre-trained language model semantic vector analysis. *Applied Soft Computing*, *169*, 112551. https://doi.org/10.1016/j.asoc.2024.112551

Ji, X., Wang, J., & Yan, Z. (2021). A stock price prediction method based on deep learning technology. *International Journal of Crowd Science*, *5*(1), 55–72. https://doi.org/10.1108/IJCS-05-2020-0012

Kang, W., Yuan, X., Zhang, X., Chen, Y., & Li, J. (2024). ChatGPT-based Sentiment Analysis and Risk Prediction in the Bitcoin Market. *Procedia Computer Science*, *242*, 211–218. https://doi.org/10.1016/j.procs.2024.08.258

Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. *Finance Research Letters*, *62*, 105227. https://doi.org/10.1016/j.frl.2024.105227

Ko, H., & Lee, J. (2024). Can ChatGPT improve investment decisions? From a portfolio management perspective. *Finance Research Letters*, *64*, 105433. https://doi.org/10.1016/j.frl.2024.105433

Lee, J., Stevens, N., Han, S. C., & Song, M. (2025). A Survey of Large Language Models in Finance (FinLLMs) [arXiv:2402.02315 [cs]]. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-024-10495-6

Li, J., Lei, Y., Bian, Y., Cheng, D., Ding, Z., & Jiang, C. (2024). RA-CFGPT: Chinese financial assistant with retrieval-augmented large language model. *Frontiers of Computer Science*, *18*(5), 185350. https://doi.org/10.1007/s11704-024-31018-5

Ma, F., Lyu, Z., & Li, H. (2024). Can ChatGPT predict Chinese equity premiums? *Finance Research Letters*, *65*, 105631. https://doi.org/10.1016/j.frl.2024.105631

microsoft. (2020). Proceso de ciencia de datos en equipo (TDSP). https://learn.microsoft.com/es-es/azure/architecture/data-science-process/overview

Mukherjee, R., Bohra, A., Banerjee, A., Sharma, S., Hegde, M., Shaikh, A., Shrivastava, S., Dasgupta, K., Ganguly, N., Ghosh, S., & Goyal, P. (2022). ECTSum: A New Benchmark Dataset For Bullet Point Summarization of Long Earnings Call Transcripts. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10893–10906. https://doi.org/10.18653/v1/2022.emnlp-main.748

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., & Awadallah, A. (2023, June). Orca: Progressive Learning from Complex Explanation Traces of GPT-4 [arXiv:2306.02707 [cs]]. Retrieved November 11, 2024, from http://arxiv.org/abs/2306.02707

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052

Onan, A., & Dursun, E. D. (2024). Benchmarking Retrieval Augmented Generation in Quantitative Finance [Series Title: Lecture Notes in Networks and Systems]. In C. Kahraman, S. Cevik Onar, S. Cebi, B. Oztaysi, A. C. Tolga, & I. Ucal Sari (Eds.), *Intelligent and Fuzzy Systems* (pp. 64–74, Vol. 1089). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-67195-1_9

Panickssery, A., Bowman, S. R., & Feng, S. (2024, April). LLM Evaluators Recognize and Favor Their Own Generations [arXiv:2404.13076 [cs]]. https://doi.org/10.48550/arXiv.2404.13076

Paul, J., Roy, A., Mitra, A., & Sil, J. (2024). HyV-Summ: Social media video summarization on custom dataset using hybrid techniques. *Neurocomputing*, 128852. https://doi.org/10.1016/j.neucom.2024.128852

Pelster, M., & Val, J. (2024). Can ChatGPT assist in picking stocks? *Finance Research Letters*, *59*, 104786. https://doi.org/10.1016/j.frl.2023.104786

Pieper, T., Ballout, M., Krumnack, U., Heidemann, G., & Kühnberger, K.-U. (2024, September). Enhancing SLM via ChatGPT and Dataset Augmentation [arXiv:2409.12599 [cs]]. Retrieved November 11, 2024, from http://arxiv.org/abs/2409.12599

Pinto, N., Da Silva Figueiredo, L., & Garcia, A. C. (2021). Automatic Prediction of Stock Market Behavior Based on Time Series, Text Mining and Sentiment Analysis: A Systematic Review. *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 1203–1208. https://doi.org/10.1109/CSCWD49262.2021.9437732

Ren, R., Wu, D. D., & Liu, T. (2019). Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine. *IEEE Systems Journal*, *13*(1), 760–770. https://doi.org/10.1109/JSYST.2018.2794462

Saleh, M. E., Wazery, Y. M., & Ali, A. A. (2024). A systematic literature review of deep learning-based text summarization: Techniques, input representation, training strategies, mechanisms, datasets, evaluation, and challenges. *Expert Systems with Applications*, *252*, 124153. https://doi.org/10.1016/j.eswa.2024.124153

Shakil, H., Farooq, A., & Kalita, J. (2024). Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing*, *603*, 128255. https://doi.org/10.1016/j.neucom.2024.128255

Shao, Z., Yao, X., Chen, F., Wang, Z., & Gao, J. (2024). Revisiting time-varying dynamics in stock market forecasting: A Multi-source sentiment analysis approach with large language model. *Decision Support Systems*, 114362. https://doi.org/10.1016/j.dss.2024.114362

Sun, K., & Wang, R. (2024, July). Textual Similarity as a Key Metric in Machine Translation Quality Estimation [arXiv:2406.07440 [cs]]. https://doi.org/10.48550/arXiv.2406.07440

Syamala Rao M., P. N. V., Kumar, N. S., Kollapudi, P., Babu, C. M., & Tara, S. (2023). A review of the techniques of fundamental and technical stock analysis, 030062. https://doi.org/10.1063/5.0125210

Turc, I., Chang, M.-W., Lee, K., & Toutanova, K. (2019, September). Well-Read Students Learn Better: On the Importance of Pre-training Compact Models [arXiv:1908.08962 [cs]]. Retrieved November 11, 2024, from http://arxiv.org/abs/1908.08962

Udagawa, T., Trivedi, A., Merler, M., & Bhattacharjee, B. (2023). A Comparative Analysis of Task-Agnostic Distillation Methods for Compressing Transformer Language Models. *Proceedings of the 2023 Conference on Empirical Methods*

in Natural Language Processing: Industry Track, 20–31. https://doi.org/10.18653/v1/2023.emnlp-industry.3

Xu, Y., Xu, R., Iter, D., Liu, Y., Wang, S., Zhu, C., & Zeng, M. (2023, May). InheritSumm: A General, Versatile and Compact Summarizer by Distilling from GPT [arXiv:2305.13083 [cs]]. Retrieved November 11, 2024, from http://arxiv.org/abs/2305.13083

Zhang, X., Li, Y., Wang, J., Sun, B., Ma, W., Sun, P., & Zhang, M. (2024, June). Large Language Models as Evaluators for Recommendation Explanations [arXiv:2406.03248 [cs]]. https://doi.org/10.48550/arXiv.2406.03248

A Anexos

A.1 Repositorio con el proceso de ingesta de transcripciones y metadatos desde YouTube

Repositorio en GitHub: https://github.com/AndresR2909/youtube-ingest/blob/master/README.md

A.2 Repositorio de creación de conjunto de datos, ajuste fino, evaluación y despliegue

Repositorio en GitHub: https://github.com/AndresR2909/poc-summary/tree/main

A.3 Conjunto de datos de entrenamiento y evaluación

Disponible en Huggingface: https://huggingface.co/datasets/AndresR2909/youtube_transcriptions_summaries_2025_gpt4.1.

A.4 Costos de generación del conjunto de datos de entrenamiento con API OpenAI (gpt4.1)

En la Figura 21 se observan los costos de generación de conjunto de datos de entrenamiento.

A.5 Modelo seleccionado guardado en HuggingFace

- Modelo 16 bits: https://huggingface.co/AndresR2909/unsloth_Meta-Llama-3.1-8B-Instruct-bnb-4bit_16bit_v2
- Modelo cuantizado (4 y 8 bits): https://huggingface.co/AndresR2909/unsloth_Meta-Llama-3.1-8B-Instruct-bnb-4bit_gguf_v2

A.6 Map reduce and iterative refinement

Referencia: https://python-langchain-com.translate.goog/docs/tutorials/summarization/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc

A.7 Reporte en wandDB de ajuste fino de modelos

https://wandb.ai/felipeandres29-universidad-eafit/finetune-Llama-3.2-3B-Instruct/reports/Comparaci-n-Finetuning-de-modelos-llama-3-1-y-3-2-usando-LoRa-y-QLoRa---VmlldzoxM
accessToken=plse4z2tpajhdga7e3qoqqtxy9u108wx49o78325kv8zlsfbtr0dfq2of4ovqv2i

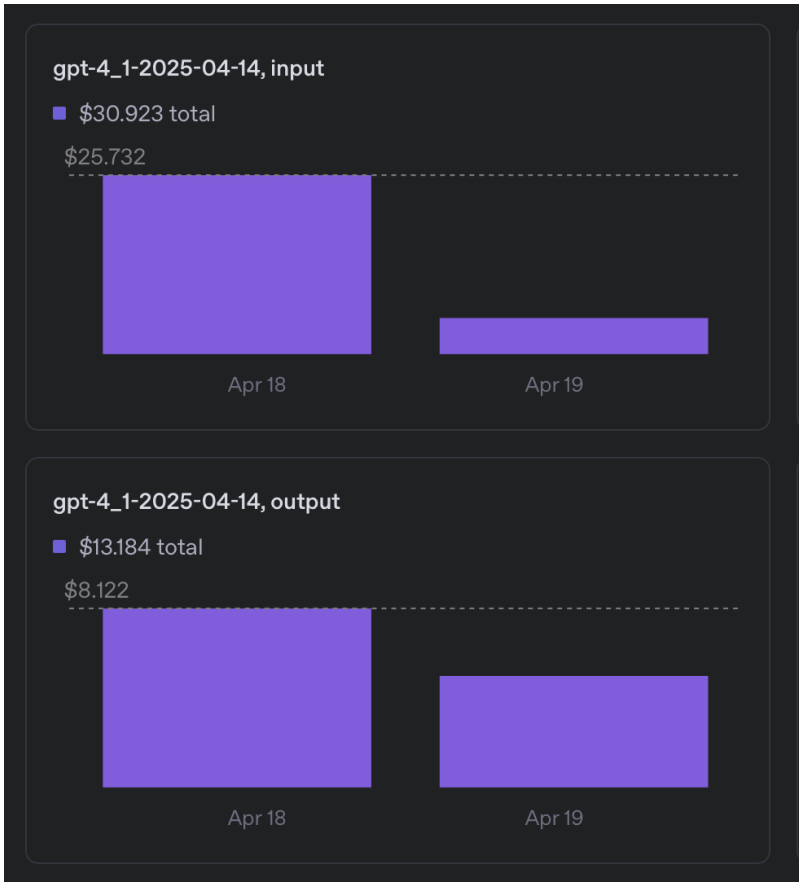

Figure 21: Costos de generación de conjunto de datos de entrenamiento.