RESPUESTAS A PREGUNTAS EN CONTRATOS DE ARRENDAMIENTO BAJO LA NORMATIVA ASC (ACCOUNTING STANDARDS CODIFICATION) 842 UTILIZANDO GRANDES MODELOS DE LENGUAJE

Answers to Questions in Lease Contracts under ASC (Accounting Standards Codification) 842 Using Large Language Models

Estudiante: David Adrián Armendáriz Peña

Proyecto de grado

Director: Tomás Olarte Hernández

Universidad EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

Maestría en Ciencia de Datos y Analítica

Medellín

2024

Tabla de contenidos

Resumen 5

Abstract 6

Descripción del proyecto 7

- Planteamiento del problema 7
- Justificación 8
- Objetivos Generales 8
- Objetivos Específicos 9

Marco teórico 10

- GAAP 10
- ASC 842 10
- Modelos de Lenguaje de Gran Escala (LLM) 11
- Embeddings 11
- Arquitectura Transformer 11
- Aumentar el conocimiento de los LLMs 11
- Full-model fine-tuning 12
- LoRA fine-tuning 12
- RAG 12
- Respuestas a preguntas 13
- Ingeniería de prompts 14
- Bases de datos vectoriales 14

Estado del arte 15

- Modelos LLM de OpenAI 15
- Modelos de embeddings de OpenAI 16
- Bases de datos vectoriales 17
- Técnicas de ingeniería de prompts 18
  - Zero-shot prompting 18
  - Few-shot prompting 18
  - Chain of Thought (CoT) 18
- Salidas estructuradas 18

Aspectos éticos 19

- ¿Para qué se van a usar los datos? 19
- ¿Cuáles son los beneficios y quién se beneficiará? 19

¿Quién estará usando los datos? 19
Mecanismos para obtener consentimiento de los propietarios de la información 19
Mecanismos para anonimizar los datos 19
Mecanismos para garantizar la seguridad de la información 19
Metodología 20
GenAI Lifecycle 20
Identificación del problema y motivación 21
Investigación de datos 23
Configuración de la consulta para obtener contratos del SEC 23
Variabilidad de los contratos 26
Preparación de los datos 27
Desarrollo 29
Interfaz de usuario 29
API de recepción y almacenamiento 31
API de procesamiento de archivos 31
API de generación de respuestas 32
Experimentos y prompts utilizados 32
Evaluación 36
Plan de evaluación 36
Procedimientos de Evaluación 37
Testeo automatizado 37
Análisis de errores 38
Reporte de resultados 38
Resultados 38
Mejoras para la pregunta 1 41
Mejoras para la pregunta 2 46
Mejoras para la pregunta 3 48
Mejoras para la pregunta 4 49
Mejoras para la pregunta 5 50
Mejoras para la pregunta 6 51
Mejoras para la pregunta 7 52
Mejoras para la pregunta 8 53
Resumen de los mejores prompts 54

Despliegue 57
Enfoque con AWS Lambda y API Gateway 57
Enfoque con Elastic Beanstalk y con una sola instancia 58
Enfoque con Elastic Beanstalk y un balanceador de carga 59
Monitoreo y mejoras 60
Trabajo futuro 61
Conclusiones 62
Referencias 65
Anexos 67

Resumen

La normativa ASC 842, parte de los GAAP (Generally Accepted Accounting Principles en inglés) en Estados Unidos, establece reglas para registrar arrendamientos en los balances financieros, mejorando la transparencia y la comparabilidad. Sin embargo, su implementación presenta retos significativos, como interpretar contratos complejos y extraer información clave, tareas que suelen realizarse de forma manual, lo que implica altos costos y errores.

Esta tesis desarrolla un sistema automatizado para responder preguntas relevantes sobre contratos de arrendamiento, utilizando Procesamiento de Lenguaje Natural, Grandes Modelos de Lenguaje y Generación Aumentada por Recuperación. El objetivo es reducir la dependencia de consultores externos al identificar información necesaria para elaborar memorandos técnicos contables de manera automática.

Se utilizó la metodología GenAI Lifecycle, que incluye la vectorización de texto mediante modelos de embeddings y el almacenamiento de datos en bases vectoriales como Pinecone. Con contratos de arrendamiento obtenidos de la SEC (Security Exchange Comission en inglés), el sistema se desarrolló para responder preguntas clave como fechas, nombres de arrendatarios, opciones de compra o renovación, alcanzando una precisión de al menos el 70%.

Los resultados demuestran que el sistema reduce significativamente los tiempos y costos asociados al análisis de contratos, mejorando la precisión en el cumplimiento de la normativa ASC 842. Este enfoque tiene implicaciones prácticas para la industria contable, al ofrecer una solución escalable que democratiza el acceso a herramientas avanzadas de inteligencia artificial, permitiendo a las empresas gestionar de manera eficiente sus procesos normativos.

Este trabajo representa un avance significativo en la integración de inteligencia artificial para resolver problemas reales en la contabilidad, fomentando la innovación en la extracción y análisis de información normativa.

Palabras clave: Grandes Modelos de Lenguaje, Generación Aumentada por Recuperación, Codificación de Normas Contables, Norma de contabilidad ASC 842, Contabilidad de Arrendamientos

Abstract

The ASC 842 standard, part of GAAP (Generally Accepted Accounting Principles) in the United States, establishes rules for recording leases in financial statements, enhancing transparency and comparability. However, its implementation poses significant challenges, such as interpreting complex contracts and extracting key information, tasks often performed manually, leading to high costs and errors.

This thesis develops an automated system to address relevant questions about lease contracts using Natural Language Processing, Large Language Models, and Retrieval Augmented Generation. The goal is to reduce reliance on external consultants by identifying the information needed to draft technical accounting memos automatically.

The GenAI Lifecycle methodology was employed, including text vectorization using embedding models and data storage in vector databases like Pinecone. Using lease contracts obtained from the Security Exchange Comission, the system was developed to answer key questions such as dates, purchase options, or renewal terms, achieving at least 70% accuracy.

The results demonstrate that the system significantly reduces the time and costs associated with contract analysis, improving the accuracy in compliance with ASC 842. This approach has practical implications for the accounting industry, offering a scalable solution that democratizes access to advanced artificial intelligence tools, enabling companies to efficiently manage their regulatory processes.

This work represents a significant step forward in integrating artificial intelligence to solve real-world accounting problems, fostering innovation in the extraction and analysis of regulatory information.

Keywords: Large Language Models, Retrieval-Augmented Generation, Accounting Standards Encoding, ASC 842 Accounting Standard, Lease Accounting

Descripción del proyecto

Planteamiento del problema

Los principios contables ayudan a que los informes financieros de una empresa cumplan con estándares regulados. En Estados Unidos, estos estándares se conocen como GAAP. Las empresas que deben cumplir con los estándares GAAP deben hacerlo en todos los informes financieros o corren el riesgo de enfrentar consecuencias significativas (Crail & Main, 2022).

La contabilidad de arrendamientos ha sido un aspecto importante en el mundo de las finanzas corporativas durante décadas. Sin embargo, con la emisión de la normativa ASC 842 por la Junta de Normas de Contabilidad Financiera (Financial Accounting Standards Board o FASB por sus siglas en inglés) en los Estados Unidos, se ha intensificado la necesidad de un enfoque más riguroso y eficiente en la gestión de los contratos de arrendamiento. Esta normativa, que las empresas públicas estaban obligadas a adoptar en 2019 y, para empresas privadas, en 2020, ha impuesto retos importantes en la forma en que las organizaciones deben contabilizar y divulgar sus arrendamientos operativos y financieros en los estados financieros (Accruent, 2023).

A medida que las empresas se adaptan a los requisitos de ASC 842, se enfrentan a desafíos considerables en la generación de memorandos que documenten adecuadamente la información clave contenida en los contratos de arrendamiento. Actualmente, este proceso se lleva a cabo de manera manual, lo que conlleva costos considerables en términos de tiempo y recursos humanos.

La generación de memorandos que detallan las disposiciones clave de los contratos de arrendamiento implica una revisión minuciosa de documentos extensos y complejos. Esta revisión manual no solo es propensa a errores humanos, sino que también es inherentemente lenta y laboriosa. Los auditores y profesionales financieros se ven obligados a dedicar una cantidad significativa de tiempo y esfuerzo a la extracción manual de información relevante, lo que retrasa el proceso de revisión y agrega costos adicionales a las empresas.

Además, la complejidad inherente de los contratos de arrendamiento, con cláusulas y disposiciones variadas, dificulta aún más el proceso de identificación de información relevante. La falta de una metodología estandarizada y eficiente para la extracción de datos relevantes aumenta la probabilidad de errores y omisiones, lo que a su vez puede conducir a una divulgación inadecuada de información financiera y, en última instancia, a posibles sanciones regulatorias.

En este contexto, surge la necesidad de desarrollar enfoques innovadores que permitan una extracción de información más eficiente y precisa de los contratos de arrendamiento, con el fin de agilizar el proceso de generación de memorandos para su revisión por parte de auditores en los Estados Unidos. La aplicación de tecnologías de lenguaje natural, en particular, los Grandes Modelos de Lenguaje (Large Language Models o LLM por sus siglas en inglés) y técnicas como Generación Aumentada por Recuperación (Retrieval Augmented Generation o RAG por sus siglas en inglés), ofrece un potencial considerable para abordar estos desafíos y mejorar significativamente la eficiencia y precisión en la generación de memorandos para la contabilidad de arrendamientos.

En este sentido, esta tesis se propone explorar y evaluar el uso de LLMs, en combinación con la metodología RAG, para la extracción automatizada de información de contratos de arrendamiento

sujetos a ASC 842. Se busca no solo identificar las variables clave relevantes para la contabilidad de arrendamientos, sino también desarrollar un enfoque sistemático y eficiente para su extracción y documentación. Este enfoque tiene el potencial de reducir significativamente los costos y el tiempo asociados con la generación de memorandos, al tiempo que mejora la precisión y la exhaustividad de la información recopilada.

En resumen, la complejidad y el volumen de los contratos de arrendamiento representan un desafío significativo para las empresas y los auditores en los Estados Unidos. La falta de eficiencia en la generación de memorandos manualmente, debido a la extracción de información clave, agrega una carga financiera y operativa adicional a las organizaciones. Por lo tanto, es imperativo desarrollar soluciones innovadoras que aprovechen el poder de la inteligencia artificial y el procesamiento de lenguaje natural para abordar estos desafíos y mejorar la eficiencia y precisión en la gestión de la contabilidad de arrendamientos.

Justificación

La justificación de esta tesis radica en su potencial para transformar fundamentalmente la forma en que las empresas gestionan la contabilidad de arrendamientos, particularmente en el contexto de la normativa ASC 842. Al desarrollar un enfoque automatizado basado en LLMs y RAG para la extracción de información de contratos de arrendamiento, este proyecto tiene el poder de reducir significativamente los costos y el tiempo asociados con la generación de memorandos.

Además, al estandarizar y mejorar la precisión en el proceso de extracción de datos clave, este enfoque puede eliminar la necesidad de recurrir a consultores externos que tradicionalmente cobran tarifas elevadas por servicios similares. Al hacer que este proceso sea más accesible y económico, este proyecto puede democratizar el acceso a herramientas avanzadas de contabilidad de arrendamientos, permitiendo que empresas de todos los tamaños mejoren su cumplimiento normativo y optimicen sus operaciones financieras. Por ende, este proyecto tiene el potencial de generar un impacto significativo en la industria al mejorar la eficiencia, la precisión y la accesibilidad de la gestión de la contabilidad de arrendamientos en el contexto de ASC 842.

Objetivos Generales

Implementación de un sistema de detección de entidades sobre contratos de arrendamiento usable por profesionales y consultores de contabilidad que identifique 8 de las principales entidades relevantes con al menos un 70% precisión (accuracy).

El objetivo principal de esta tesis es desarrollar e implementar un sistema de detección de entidades específicamente diseñado para contratos de arrendamiento, con el propósito de facilitar el trabajo de profesionales y consultores de contabilidad. La implementación de este sistema busca cubrir una necesidad crucial en el ámbito contable, donde la identificación precisa de entidades en contratos de arrendamiento es fundamental para la gestión adecuada de activos financieros y la elaboración de informes precisos. La característica distintiva de este sistema radica en su enfoque hacia la usabilidad, definida en este contexto como la capacidad de identificar 8 de las principales entidades relevantes comunes a todos los contratos de arrendamiento con al menos un 70% de precisión (accuracy).

Esta definición establece un estándar claro y medible para la efectividad del sistema, asegurando que sea práctico y funcional para los usuarios finales, y que contribuya significativamente a mejorar la eficiencia y precisión en el proceso de análisis contable.

La definición de una precisión del 70% se basa en la decisión del negocio. Una precisión menor al 70% generaría demasiados errores, lo que haría que los usuarios pierdan confianza en el sistema y requieran hacer todo manualmente. Una precisión mayor al 80%-90% podría ser innecesariamente costosa, ya que los contadores aún deben verificar ciertos datos críticos. El umbral del 70% permite que el sistema sea confiable para el análisis preliminar, reduciendo la carga de trabajo sin comprometer la calidad de los informes contables.

Objetivos Específicos

1. **Implementación del sistema para vectorizar y almacenar el texto de los contratos que por ahora solo están en formato PDF**
   Se implementará un sistema para la gestión de documentos PDF, permitiendo a los usuarios cargar archivos y almacenar la información en AWS S3 y en una base de datos vectorial para su posterior procesamiento y recuperación de información. Esto implica el desarrollo de un microservicio en Python para procesar los documentos cargados previamente en AWS S3.
2. **Implementación de un sistema de respuestas a preguntas cerradas**
   Se implementará un sistema que tome los vectores guardados en una base de datos vectorial, los recupere y con esa información vectorizada pueda responder preguntas sobre el PDF.
3. **Comparación de distintos grandes modelos de lenguaje de OpenAI**
   Se comparará distintos modelos de pago de OpenAI con el fin de identificar la mejor solución para el contexto de la extracción de información de contratos de arrendamiento y entender el costo beneficio de estos.
4. **Implementación de una interfaz gráfica para elección de respuestas**
   Se implementará una interfaz intuitiva y fácil de usar para que los usuarios interactúen con el sistema, permitiéndoles seleccionar las respuestas generadas por el modelo RAG que consideren más adecuadas en función del contexto de la pregunta planteada. Esta interfaz gráfica se desarrollará utilizando el framework de React.
5. **Implementación de DevSecOps para el sistema**
   El despliegue se hará en servidores en la nube de AWS, utilizando Elastic Beanstalk y será conectado con un backend existente que utiliza el framework Ruby on Rails, asegurando una integración fluida entre los diferentes componentes del sistema. El sistema debe ser seguro, por lo que se buscará que tenga un certificado SSL para soportar el protocolo HTTPS.
   Se implementará Encriptación en Reposo (Encryption at Rest en inglés) de los documentos, ya que los contratos de arrendamiento son información importante y confidencial de una empresa, por lo que es imperativo la encriptación de estos en el almacenamiento. También, en la API se implementará un mecanismo de autenticación, para que solamente el backend pueda acceder a su servicio. Debido a que la inteligencia artificial puede alucinar y queremos evitar dar respuestas incorrectas al usuario también se implementará un sistema de monitoreo y retroalimentación para recoger la calidad de la respuesta del usuario.

Marco teórico

GAAP

GAAP es un conjunto de normas y directrices contables detalladas diseñadas para asegurar que las empresas que cotizan en bolsa en Estados Unidos reporten información financiera de manera clara y coherente. Siguiendo los procedimientos de GAAP, una empresa puede generar un informe financiero comparable con el de otras empresas del mismo sector. Esto facilita a inversores, acreedores y otras partes interesadas investigar y evaluar financieramente una empresa u organización de manera eficiente. Con GAAP, incluso aspectos específicos como la preparación de impuestos y la declaración de activos o pasivos se presentan de forma estandarizada (Crail & Main, 2022).

GAAP es administrado y publicado por FASB, que actualiza periódicamente los principios y normas. Es el equivalente estadounidense de las Normas Internacionales de Información Financiera (NIIF). Aunque solo las empresas reguladas y que cotizan en bolsa están legalmente obligadas a seguir GAAP, algunas empresas privadas también deciden adherirse a estos estándares en sus estados financieros (Crail & Main, 2022).

Si se encuentra que una empresa está violando los principios de GAAP, puede enfrentar numerosas consecuencias. Estas pueden incluir grandes multas monetarias, impactos negativos significativos en su credibilidad y problemas financieros internos debido a una contabilidad incorrecta. Siempre es más ventajoso cumplir con las directrices de GAAP desde el principio, ya que no hacerlo puede resultar en la pérdida de posibles inversores y oportunidades al no mantener un trabajo de alta calidad (Crail & Main, 2022).

ASC 842

ASC 842 es la nueva norma de contabilidad de arrendamientos emitida por FASB. Las empresas públicas tuvieron que adoptarla en 2019 y las empresas privadas en 2020. ASC 842 exige el seguimiento y la divulgación de todos los activos arrendados por una empresa, sustituyendo a la anterior norma de arrendamientos ASC 840 (Accruent, 2023).

La FASB emitió la ASU 2016-02 (Accounting Standard Update por sus siglas en inglés) en febrero de 2016. Aunque el proyecto comenzó como un proyecto conjunto con el IASB (International Accounting Standards Board por sus siglas en inglés), los consejos divergieron en algunas áreas clave. Por ejemplo, los consejos no estuvieron de acuerdo en si todos los arrendamientos deberían contabilizarse utilizando el mismo modelo. Después de varias discusiones, el IASB decidió que los arrendatarios deberían aplicar un único modelo a todos los arrendamientos, lo cual se refleja en la NIIF 16, Arrendamientos, publicada en enero de 2016. El FASB decidió que los arrendatarios deberían aplicar un modelo dual. Bajo el modelo del FASB, los arrendatarios clasificarán un arrendamiento como un arrendamiento financiero o un arrendamiento operativo, mientras que un

arrendador clasificará un arrendamiento como un arrendamiento de tipo de venta, de financiamiento directo u operativo (PwC, 2024).

Modelos de Lenguaje de Gran Escala (LLM)

Los modelos de lenguaje de gran escala (LLM) tienen sus raíces en los experimentos con redes neuronales iniciados en la década de 1950, con avances significativos como Eliza en los años 60, las redes neuronales LSTM en 1997 y el desarrollo de herramientas como CoreNLP en 2010. La evolución clave ocurrió con la introducción de los Transformers en 2017 y el lanzamiento de BERT en 2019, lo que revolucionó el procesamiento del lenguaje natural al mejorar la capacidad de los modelos para comprender el contexto y manejar tareas lingüísticas complejas (Toloka, 2023).

Embeddings

Los embeddings son representaciones vectoriales densas que capturan relaciones semánticas y contextuales en el procesamiento del lenguaje natural. Facilitan la conversión de texto en datos comprensibles para modelos de aprendizaje automático, permitiendo una mejor interpretación del significado y las relaciones entre palabras. Su integración con mecanismos de atención y codificación posicional ha optimizado la eficiencia del entrenamiento y el rendimiento en tareas como la traducción automática y el análisis de sentimientos (Worth, 2023).

Arquitectura Transformer

Los Transformers, propuestos por Google en 2017, superaron a las redes neuronales recurrentes al mejorar la eficiencia en la traducción automática y otras tareas de NLP. Su arquitectura se basa en encoders y decoders con mecanismos de autoatención, que permiten evaluar simultáneamente todas las palabras de una secuencia y capturar dependencias a largo plazo. Modelos como GPT y BERT han demostrado el poder de los Transformers al eliminar la necesidad de entrenamientos específicos para cada tarea y establecer nuevos estándares en el procesamiento del lenguaje natural (Vaswani et al., 2017).

Aumentar el conocimiento de los LLMs

La problemática de los LLMs radica en su falta de conocimiento específico de dominio. Estos modelos se entrenan en grandes cantidades de datos lingüísticos generales, pero no tienen acceso directo a información confidencial o documentos específicos de una empresa o un caso de uso en particular. Por lo tanto, cuando se les consulta sobre temas específicos de una organización, pueden generar respuestas que carecen de detalles precisos o que incluso son incorrectas. Para abordar esta limitación, se requiere un enfoque de adaptación o ampliación que incorpore datos específicos para mejorar la relevancia y la exactitud de las respuestas. Este proceso debe manejarse cuidadosamente para proteger la privacidad y la seguridad de la información.

En el estado del arte, el proceso de aumentar el conocimiento de modelos existentes con datos adicionales puede lograrse mediante tres métodos principales: full-model fine-tuning, LoRA fine-tuning y RAG.

El full-model fine-tuning implica el ajuste de todos los pesos en un modelo pre-entrenado utilizando datos específicos de la tarea a realizar. A pesar de su efectividad, no es factible en la práctica para

modelos grandes, debido a los costos asociados con el reentrenamiento del modelo y su tiempo de ejecución (Sun et al., 2023).

Full-model fine-tuning

El full-model fine-tuning implica el ajuste de todos los pesos en un modelo pre-entrenado utilizando datos específicos de la tarea a realizar. A pesar de su efectividad, no es factible en la práctica para modelos grandes, debido a los costos asociados con el reentrenamiento del modelo y su tiempo de ejecución (Sun et al., 2023).

LoRA fine-tuning

Un paradigma importante del procesamiento del lenguaje natural consiste en el preentrenamiento a gran escala con datos de dominio general y la adaptación a tareas o dominios específicos. A medida que se pre-entrenan modelos más grandes, el ajuste fino completo, que vuelve a entrenar todos los parámetros del modelo, se vuelve menos factible. LoRA (Low-Rank Adaptation) congela los pesos del modelo pre-entrenado e introduce matrices de descomposición de rango entrenables en cada capa de la arquitectura Transformer, lo que reduce significativamente la cantidad de parámetros que hay que entrenar (Hu et al., 2021).

RAG

Tanto el full-model fine-tuning como el LoRA fine-tuning implican un reentrenamiento. La técnica de RAG ayuda a aumentar información adicional, sin reentrenar el modelo. El RAG toma datos adicionales y los deposita en una base de datos vectorial después de transformarlos en embeddings. Esto se hace solo una vez. Si los datos están evolucionando, simplemente sigue depositando los embeddings en la base de datos vectorial. No es necesario repetir esto nuevamente para todos los datos. Luego, se utiliza el mismo modelo de embeddings para transformar la consulta del usuario a un vector. Después, encuentra los vecinos más cercanos en la base de datos vectorial a la consulta vectorizada. Por último, se proporciona la consulta original y los documentos recuperados (para obtener más contexto) al LLM para obtener una respuesta (Lewis et al., 2020).

Según Merrit (2023), este proceso se puede visualizar de manera más clara con el siguiente diagrama:

Figura 1 Ejemplo de utilización de RAG usando una base de datos vectorial (Jing et al., 2024)

RAG ha surgido como una solución eficaz al incorporar conocimientos de fuentes externas. Esto mejora la precisión y la credibilidad del texto generado, particularmente para tareas que requieren conocimiento específico ya que permite actualizaciones de conocimiento continuas e integración de información específica del dominio.

Según Amazon (2024b) hay muchas ventajas para que una empresa implemente RAG:

- Implementación rentable
    - El desarrollo de chatbots generalmente comienza utilizando un modelo fundacional. Los modelos fundacionales (Foundational Models o FMs por sus siglas en inglés) son

LLMs accesibles a través de APIs (Application Programming Interface en inglés) y entrenados con un amplio conjunto de datos generalizados y no etiquetados. Los costos computacionales y financieros de volver a entrenar los FMs para información específica de una organización o dominio son altos. RAG es un enfoque más rentable para introducir nuevos datos en el LLM y hace que la tecnología de inteligencia artificial generativa sea más accesible.

- Información actual
    - Incluso si las fuentes de datos originales de entrenamiento para un LLM son adecuadas para el caso de uso, es importante mantener la relevancia de la información. RAG permite a los desarrolladores proporcionar datos actualizados a los FMs. Incluso, se puede usar RAG para conectar el LLM directamente a flujos de redes sociales en vivo, sitios de noticias u otras fuentes de información que se actualizan frecuentemente. Así, el LLM puede proporcionar la información más reciente a los usuarios.
- Mayor confianza del usuario
    - El RAG permite que el LLM presente información precisa con las fuentes. Es decir, el texto puede incluir citas o referencias a las fuentes. Los usuarios también pueden consultar los documentos fuente ellos mismos si necesitan más detalles. Esto puede aumentar la confianza y la seguridad en la solución de IA generativa.
- Mayor control para los desarrolladores
    - Con RAG, los desarrolladores pueden probar y mejorar sus aplicaciones de chat de manera más eficiente. Pueden controlar y cambiar las fuentes de información del LLM para adaptarse a requisitos cambiantes o usos multifuncionales. Los desarrolladores también pueden restringir la recuperación de información sensible a diferentes niveles de autorización y asegurar que el LLM genere respuestas apropiadas. Además, pueden solucionar problemas y realizar correcciones si el LLM referencia fuentes de información incorrectas para preguntas específicas. Las organizaciones pueden implementar tecnología de IA generativa con mayor confianza para una gama más amplia de aplicaciones.

También hay muchos problemas con RAG. Por ejemplo, RAG implica una coincidencia en términos de similitud entre la consulta y los vectores depositados (embeddings). Sin embargo, las preguntas son estructuralmente muy diferentes de las respuestas. Por ende, usualmente se recuperan muchos documentos irrelevantes (Gao et al., 2023).

Otra problemática, es que RAG puede enfrentar el problema de alucinación, donde generan contenido que no está respaldado por datos reales. Esto ocurre especialmente cuando se manejan consultas fuera del alcance de los datos de entrenamiento o cuando se requiere información actualizada. La generación precisa y confiable sigue siendo un desafío (Gao et al., 2023).

Respuestas a preguntas

La Respuesta a Preguntas (Question Answering o QA, por sus siglas en inglés) es una de las tareas más importantes del NLP. Su objetivo es utilizar tecnologías de NLP para generar una respuesta correspondiente a una pregunta dada, basada en un enorme corpus no estructurado (Wang, 2022). QA es un tema de investigación que se propuso hace medio siglo. Los sistemas tradicionales de QA

integran algunas técnicas de recuperación de información (como RAG) para encontrar respuestas (Wang, 2022).

Según Hugging Face (2024), QA tiene varias variantes:

- **QA extractiva:** El modelo extrae la respuesta de un contexto. El contexto aquí podría ser un texto proporcionado, una tabla o incluso HTML. Esto generalmente se resuelve con modelos tipo BERT.
- **QA generativa abierta:** El modelo genera texto libre directamente basado en el contexto. Puedes aprender más sobre la tarea de generación de texto en su página.
- **QA generativa cerrada:** En este caso, no se proporciona ningún contexto. La respuesta es completamente generada por un modelo.

Según Hugging Face (2024), para evaluar qué tan eficiente es Q&A, existen dos métricas:

- **Exact match:** Exact match es una métrica basada en la coincidencia estricta de caracteres entre la respuesta predicha y la respuesta correcta. Para las respuestas predichas correctamente, el Exact Match será 1. Incluso si solo un carácter es diferente, el Exact Match será 0.
- **F1:** La métrica F1-Score es útil si valoramos tanto los falsos positivos como los falsos negativos de manera igual. El F1-Score se calcula sobre cada palabra en la secuencia predicha en comparación con la respuesta correcta.

Ingeniería de prompts

La ingeniería de prompts es una disciplina emergente que optimiza las interacciones con modelos de inteligencia artificial generativa. Un prompt es la interfaz textual que permite a los usuarios guiar la salida de un modelo, ya sea en generación de imágenes (DALLE-3, Midjourney) o en modelos de lenguaje (GPT-4, Gemini). Su diseño puede incluir instrucciones, datos y ejemplos para mejorar la precisión de las respuestas (Amatriain, 2024).

El desarrollo de prompts efectivos requiere comprender las capacidades y limitaciones del modelo, además del contexto de uso. Más que una simple instrucción, implica un enfoque metódico que puede incluir plantillas dinámicas para generar respuestas personalizadas.

Este proceso es iterativo y se asemeja a prácticas de ingeniería de software como el control de versiones y las pruebas de regresión. Su crecimiento podría transformar el aprendizaje automático, integrando nuevas metodologías para optimizar la interacción con modelos de IA a gran escala (Amatriain, 2024).

Bases de datos vectoriales

Si bien los modelos de lenguaje a gran escala (LLMs) son conceptos relativamente nuevos, los sistemas de gestión de bases de datos (DBMS) han sido ampliamente desarrollados y aplicados en diversas áreas durante los últimos 60 años, siendo reconocidos por su estabilidad y universalidad en el manejo de datos estructurados con formatos fijos, optimizados para el almacenamiento en computadoras. Sin embargo, el desarrollo y la amplia aplicación de modelos de aprendizaje profundo, como las redes neuronales convolucionales y los transformadores, han permitido la

incrustación de datos no estructurados y multimodales, como imágenes y texto, en representaciones vectoriales de longitud fija. Estas representaciones capturan características semánticas de alta dimensión de los datos originales, donde las similitudes semánticas se reflejan en las distancias entre los vectores, lo que ha generado la necesidad de un nuevo tipo de DBMS diseñado específicamente para manejar operaciones con datos vectoriales, en particular la búsqueda y el almacenamiento de vectores (Jing et al., 2024).

Por otro lado, a diferencia de los sistemas de gestión de bases de datos tradicionales (DBMS), que buscan valores exactos dentro de las bases de datos, las bases de datos vectoriales dependen en gran medida de la búsqueda de vecinos más cercanos aproximados (ANN, por sus siglas en inglés) para vectores. Esta técnica busca los k vecinos más cercanos de manera aproximada dentro del espacio de datos vectoriales de alta dimensión, sin requerir coincidencias exactas (Jing et al., 2024).

La escalabilidad y el rendimiento son otras características importantes de las bases de datos vectoriales. Están diseñadas para escalar horizontalmente, lo que significa que pueden manejar eficientemente grandes cantidades de datos sin comprometer el rendimiento. Esta capacidad es esencial para aplicaciones que requieren respuestas rápidas a consultas complejas.

La importancia de las bases de datos vectoriales se manifiesta en su aplicación en diversos campos. En el contexto actual de la inteligencia artificial generativa y el procesamiento del lenguaje natural, estas bases son fundamentales. Su capacidad para realizar búsquedas semánticas rápidas las convierte en herramientas esenciales para sistemas de reconocimiento de imágenes, donde permiten identificar y clasificar imágenes basándose en características visuales, sistemas de recomendación, que facilitan la recomendación de productos o contenidos al encontrar similitudes entre diferentes elementos y análisis contextual, que ayudan a los modelos de IA a comprender mejor el contexto y las relaciones entre distintos tipos de datos (Amazon, 2024a).

Estado del arte

Modelos LLM de OpenAI

OpenAI ha desarrollado una serie de modelos de lenguaje que han revolucionado la inteligencia artificial generativa, destacando en la comprensión y generación de texto. Entre estos modelos, **GPT-3.5 Turbo**, **GPT-4o**, y **GPT-4o Mini** son particularmente relevantes en el contexto actual.

GPT-3.5 Turbo

Lanzado como una mejora sobre sus predecesores, GPT-3.5 Turbo se diseñó para ser más eficiente en tareas de conversación y productividad. Sin embargo, a pesar de su popularidad, este modelo ha mostrado limitaciones en comparación con sus competidores más recientes. En diversas pruebas de rendimiento, como MMLU (Massive Multitask Language Understanding), GPT-3.5 Turbo no alcanzó los estándares establecidos por modelos como Google Gemini 1.5 Flash y Claude 3 Haiku, lo que ha llevado a una percepción de que es inferior en términos de capacidades (OpenAI, 2024a).

GPT-4o

Lanzado en mayo de 2024, GPT-4o representa una evolución significativa respecto a GPT-4. Este modelo no solo mejora la interacción textual, sino que también introduce capacidades multimodales, permitiendo la entrada y salida de texto e imágenes. Según OpenAI, GPT-4o está diseñado para ser más accesible, incluso para usuarios sin suscripción paga, lo que amplía su uso potencial. Este modelo ha sido elogiado por su capacidad para realizar razonamientos complejos y ha demostrado un rendimiento superior en tareas académicas y de codificación en comparación con versiones anteriores (OpenAI, 2024a).

GPT-4o Mini

El modelo **GPT-4o Mini** se lanzó como una alternativa más económica y eficiente, superando a GPT-3.5 Turbo en múltiples benchmarks académicos. Con una ventana de contexto ampliada a 128k y costos significativamente reducidos (15 centavos por millón de fichas de entrada), este modelo se posiciona como una opción atractiva para desarrolladores que buscan un alto rendimiento a un costo menor. En pruebas específicas, como MMLU, GPT-4o Mini logró un 82% de precisión, superando a su predecesor por un margen considerable (OpenAI, 2024a).

A medida que OpenAI avanza hacia el desarrollo de nuevos modelos como **o1**, que promete capacidades aún más avanzadas en razonamiento y resolución de problemas complejos, el enfoque parece estar en mejorar la eficiencia y la accesibilidad. Sin embargo, los desafíos persisten, incluyendo la necesidad de mayor potencia computacional y la gestión ética del uso de estas tecnologías avanzadas

Modelos de embeddings de OpenAI

Los modelos de embeddings han revolucionado el NLP, permitiendo representar texto en forma de vectores numéricos que facilitan diversas tareas como la clasificación, la búsqueda y la generación de texto. En este contexto, se destacan tres modelos recientes de OpenAI: text-embedding-ada-002, text-embedding-3-small, y text-embedding-3-large.

text-embedding-ada-002

Lanzado en diciembre de 2022, text-embedding-ada-002 fue diseñado para unificar varias capacidades de búsqueda y similitud textual bajo un solo modelo. Este modelo ha demostrado ser significativamente más eficiente y efectivo que sus predecesores, logrando un rendimiento superior en tareas de búsqueda y clasificación. En términos de rendimiento, este modelo alcanzó un promedio del 61% en benchmarks de tareas en inglés (MTEB) y un 31.4% en tareas multilingües (MIRACL)

text-embedding-3-small

Según OpenAI, el modelo text-embedding-3-small representa un avance significativo en eficiencia y rendimiento en comparación con su predecesor, text-embedding-ada-002, lanzado en diciembre de 2022.

En términos de rendimiento, se ha observado una mejora notable en los resultados de benchmarks. En un benchmark comúnmente utilizado para la recuperación multilingüe, MIRACL, el puntaje promedio de text-embedding-3-small ha aumentado del 31.4% al 44.0%. Además, en el benchmark MTEB, que evalúa tareas en inglés, el puntaje promedio ha incrementado de 61.0% a 62.3%. Estas

cifras demuestran que text-embedding-3-small proporciona una mayor capacidad de recuperación y relevancia en las tareas de procesamiento de lenguaje natural (OpenAI, 2024c).

Desde la perspectiva económica, text-embedding-3-small también presenta una ventaja significativa, ya que es considerablemente más eficiente que text-embedding-ada-002. Esto se refleja en una reducción de precios, donde el costo por cada 1,000 tokens ha disminuido de $0.0001 a $0.00002, lo que representa una reducción de cinco veces en el costo. Esta disminución en el precio, junto con el aumento en el rendimiento, hace que text-embedding-3-small sea una opción más atractiva para aplicaciones de procesamiento de lenguaje natural (OpenAI, 2024c).

text-embedding-3-large

Por otro lado, el modelo text-embedding-3-large se posiciona como el mejor modelo de mejor rendimiento hasta la fecha. Al comparar text-embedding-ada-002 con text-embedding-3-large, se observan mejoras significativas en los resultados de benchmarks clave. En el benchmark MIRACL, que evalúa la recuperación multilingüe, el puntaje promedio ha aumentado del 31.4% al 54.9%. Asimismo, en el benchmark MTEB, que se centra en tareas en inglés, el puntaje promedio ha incrementado de 61.0% a 64.6%.

Estas cifras evidencian que text-embedding-3-large no solo supera a text-embedding-ada-002 en términos de rendimiento, sino que también ofrece una capacidad superior para abordar tareas complejas de procesamiento de lenguaje natural. Este avance en el rendimiento resalta la efectividad de text-embedding-3-large como una solución robusta y confiable para aplicaciones que requieren una comprensión profunda del lenguaje.

Ambos modelos, text-embedding-3-small y text-embedding-3-large, superan a text-embedding-ada-002 en términos de rendimiento, adaptabilidad y eficiencia de costos. La disponibilidad de dos opciones permite a los desarrolladores seleccionar el modelo que mejor se adapte a sus necesidades específicas, ya sea priorizando la eficiencia y el costo con text-embedding-3-small o buscando un rendimiento óptimo en tareas más complejas con text-embedding-3-large. Por esta razón, se decidió solo usar estos dos modelos en combinación con gpt-4o-mini y gpt-4o-2024-08-06.

Bases de datos vectoriales

Hoy en día existen varias opciones de bases de datos vectoriales que destacan por su relevancia y capacidades.

- **Pinecone** es una plataforma de pago que ofrece alta escalabilidad y búsquedas semánticas con baja latencia, ideal para aplicaciones de inteligencia artificial.
- **Weaviate** es una base de datos vectorial de código abierto que también permite búsquedas semánticas eficientes y puede gestionar grandes volúmenes de datos.
- **Qdrant** es una base de datos vectorial de código abierto que se enfoca en la tecnología de búsqueda de similitud de vectores.
- **Elasticsearch** es un motor de búsqueda, pero últimamente ha incorporado capacidades vectoriales, combinando búsqueda tradicional con búsqueda semántica.

Técnicas de ingeniería de prompts

Zero-shot prompting

Los LLMs actuales, como GPT-3.5 Turbo y GPT-4, están ajustados para seguir instrucciones y han sido entrenados con grandes cantidades de datos. El entrenamiento a gran escala hace que estos modelos sean capaces de realizar algunas tareas de manera "zero-shot". Esta técnica implica que el prompt utilizado para interactuar con el modelo no contendrá ejemplos ni demostraciones. El prompt instruye directamente al modelo para que realice una tarea sin ejemplos adicionales que lo guíen (Prompt Engineering Guide, 2024c).

Few-shot prompting

Si bien los modelos de lenguaje grande demuestran capacidades notables de disparo cero, aún se quedan cortos en tareas más complejas cuando se usa la configuración de disparo cero. El few-shot prompting puede usarse como una técnica para habilitar el aprendizaje en contexto, donde proporcionamos demostraciones en el prompt para guiar al modelo hacia un mejor desempeño. Estas demostraciones actúan como un condicionamiento para los ejemplos posteriores, en los cuales nos gustaría que el modelo genere una respuesta (Prompt Engineering Guide, 2024b).

Chain of Thought (CoT)

El Chain of Thought es una técnica utilizada en modelos de lenguaje para mejorar su capacidad de razonamiento al descomponer problemas complejos en pasos secuenciales, en lugar de ofrecer respuestas directas (Wei et al., 2022).

Este enfoque permite que el modelo genere razonamientos intermedios, facilitando la solución de tareas que requieren lógica, cálculo o análisis estructurado. Por ejemplo, ante una pregunta matemática, el modelo no solo proporciona la respuesta final, sino que expone los pasos necesarios para llegar a ella, aumentando la transparencia y precisión de los resultados. La implementación de CoT puede realizarse mediante "prompting", utilizando ejemplos que incluyan cadenas de razonamiento, o a través de "fine-tuning", ajustando el modelo para incorporar este tipo de procesamiento en su entrenamiento. Esta técnica es particularmente útil en tareas complejas, ya que permite a los modelos de lenguaje emular procesos de pensamiento lógico similares a los humanos (Prompt Engineering Guide, 2024a).

Salidas estructuradas

El 6 de agosto de 2024, OpenAI anunció una nueva funcionalidad llamada **structured outputs** o **salidas estructuradas**. Hace un tiempo se presentó el modo JSON, una herramienta esencial para desarrolladores que buscan construir aplicaciones confiables utilizando los modelos de OpenAI. Si bien este modo mejora la generación de salidas JSON válidas, no garantiza que las respuestas del modelo se ajusten a un esquema específico. Las salidas estructurados en la API se aseguran de que las respuestas generadas por el modelo coincidan exactamente con los esquemas proporcionados por los desarrolladores.

La generación de datos estructurados a partir de entradas no estructuradas es un uso clave de la inteligencia artificial en las aplicaciones de hoy en día. Los desarrolladores emplean la API de

OpenAI para crear asistentes que pueden recuperar información, responder preguntas mediante llamadas a funciones, extraer datos estructurados para la entrada y construir flujos de trabajo complejos que permiten a los modelos de lenguaje realizar acciones. Tradicionalmente, se han superado estas limitaciones de los LLM mediante herramientas de código abierto, ajustes en las solicitudes y reintentos para garantizar que las salidas del modelo cumplan con los formatos necesarios. Las salidas estructurados resuelven este desafío al restringir los modelos de OpenAI a coincidir con los esquemas proporcionados por los desarrolladores (OpenAI, 2024e).

Aspectos éticos

¿Para qué se van a usar los datos?

Los datos provenientes de la SEC (Security Exchange Commission en inglés) se utilizarán para la validación de las técnicas descritas en la metodología del proyecto. Estos datos permitirán comprobar la efectividad y precisión de las técnicas desarrolladas en el contexto del proyecto.

¿Cuáles son los beneficios y quién se beneficiará?

Los principales beneficiarios de este proyecto serán los profesionales contables. La implementación de las técnicas validadas permitirá que estos profesionales puedan dedicar más tiempo a actividades de mayor valor agregado, optimizando sus procesos de trabajo y mejorando la eficiencia en sus labores cotidianas.

¿Quién estará usando los datos?

La empresa gaapRT será la encargada de utilizar los datos. Esta empresa implementará las técnicas validadas en su plataforma para proporcionar mejores servicios a sus usuarios.

Mecanismos para obtener consentimiento de los propietarios de la información

El consentimiento para utilizar los datos de los clientes de gaapRT está implícito en los términos de uso de gaapRT, los cuales los usuarios aceptan al utilizar los servicios de la empresa. Los términos de uso pueden consultarse en el siguiente enlace: https://www.gaaprt.com/terms-of-use.

Mecanismos para anonimizar los datos

Para el caso de los datos provenientes de la SEC, no será necesario aplicar mecanismos de anonimización ya que estos datos son públicos. Cabe destacar que no se usarán datos de clientes reales de gaapRT en este proyecto de grado, eliminando así la necesidad de anonimización para esos casos.

Mecanismos para garantizar la seguridad de la información

A pesar de que los datos utilizados son públicos, se implementarán medidas de seguridad para garantizar la protección de la información en un entorno productivo. En este sentido, se utilizará la

encriptación por defecto proporcionada por Amazon S3 para encriptar los archivos PDF. Esta medida asegura que los datos estén protegidos contra accesos no autorizados y mantengan su integridad.

Metodología

Se integrará la metodología GenAI Lifecycle para la ejecución del proyecto.

GenAI Lifecycle

GenAI Lifecycle describe los pasos para crear aplicaciones basadas en inteligencia artificial, como chatbots, asistentes virtuales o agentes inteligentes. GenAI se refiere a sistemas avanzados de aprendizaje automático capaces de crear contenido, como texto, imágenes e incluso código, que a menudo es indistinguible del contenido producido por humanos (Saltz, 2024).

Al igual que otros tipos de proyectos, un proyecto de GenAI comienza con la **definición del problema**, identificando el desafío o la oportunidad que la aplicación abordará. Luego, se realiza la **investigación de datos**, donde el equipo selecciona los datos para aumentar y/o entrenar el LLM. Luego, se lleva a cabo la **preparación de datos**, que estructura estos datos para una utilización óptima de la IA. Esto es seguido por el **desarrollo**, donde el equipo construye la aplicación GenAI. Después, la **evaluación** pone a prueba la fiabilidad y la facilidad de uso de la aplicación basada en IA. Si las pruebas son exitosas, se procede con el **despliegue**, donde la aplicación de IA se integra en su entorno operativo. El ciclo concluye con la **monitorización y mejora**, donde la retroalimentación continua refina la aplicación de IA, asegurando su relevancia y rendimiento en el mundo real (Saltz, 2024).

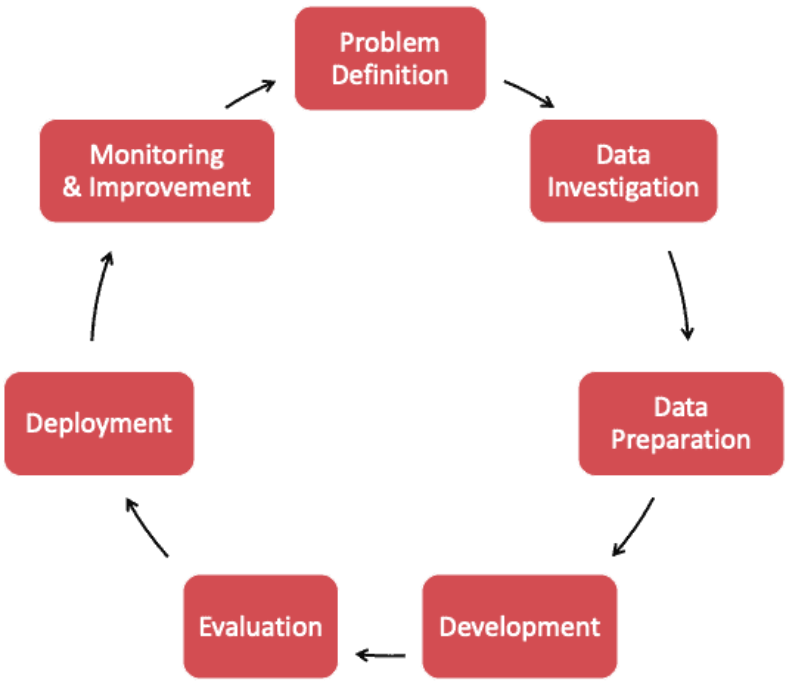

Figura 2 Ciclo de vida de GenAI (The GenAI Life Cycle, 2024).

A continuación, se describe más a fondo cada etapa del GenAI Lifecycle

- **Definición del problema**: Definir el problema, comprender su contexto empresarial y establecer objetivos claros para la solución que se va a desarrollar. Esto incluye determinar el alcance, el impacto y los resultados deseados de la aplicación.
- **Investigación de datos**: Investigar y obtener datos que puedan ser aprovechados por RAG para complementar el LLM que se está utilizando. RAG permite que el LLM acceda y utilice una amplia gama de información externa y actualizada, mejorando la capacidad del LLM para ofrecer respuestas detalladas y relevantes. Por lo tanto, esta fase se enfoca en evaluar el panorama de datos, concentrándose en la disponibilidad, relevancia y calidad de estos.
- **Preparación de datos**: Este paso implica limpiar, formatear y estructurar los datos para que sean adecuados para su uso con los modelos y tecnologías GenAI elegidos. A menudo incluye preparar los datos mediante su procesamiento y almacenamiento en una base de datos vectorial.
- **Desarrollo**: Desarrollar el agente utilizando el o los modelos de LLM adecuados, teniendo en cuenta la integración de RAG y el uso de otras técnicas de IA, como el diseño de prompts efectivos (que son instrucciones en lenguaje natural dadas o introducidas en un LLM, guiándolo para producir un resultado deseado). Esta fase también incluye, si es necesario, el fine-tuning del LLM.
- **Evaluación**: Realizar pruebas rigurosas del agente para asegurar su precisión, legibilidad, rendimiento y fiabilidad. Evaluar el agente según los criterios y objetivos predefinidos para asegurar que cumpla con los estándares y necesidades empresariales requeridos.
- **Despliegue**: Desplegar el agente en el entorno previsto, lo que incluye configurar la infraestructura necesaria. Esta configuración de infraestructura debe facilitar el alojamiento, escalado y gestión del agente, asegurando su operación fluida e integración con los sistemas existentes.
- **Monitoreo y mejora**: Implementar un monitoreo continuo de la aplicación desplegada para seguir su rendimiento, satisfacción del usuario y eficiencia operativa. Actualizar y mejorar regularmente el agente con base en los datos de rendimiento, retroalimentación de los usuarios y necesidades empresariales cambiantes.

En este proyecto se implementarán todas las fases de GenAI Lifecycle.

Identificación del problema y motivación

Hoy en día, la extracción de información relevante de contratos de arrendamiento en formato PDF se realiza de manera manual, lo que está sujeto a errores humanos. La variabilidad en la estructura de los contratos, ya que cada empresa redacta sus contratos de manera diferente, agrava este problema. No existe una forma estandarizada de extraer esta información, lo que hace que el proceso sea inconsistente y propenso a errores. Además, la necesidad de cumplir con la norma ASC 842 requiere una precisión y completitud en la extracción de datos que el proceso manual no siempre puede garantizar.

Para abordar estos desafíos, se necesitan herramientas avanzadas de procesamiento de lenguaje natural (NLP) que puedan automatizar la tarea de extracción de datos. La automatización de este

proceso no solo reduciría los errores humanos, sino que también aumentaría la eficiencia y la precisión en la generación de memorandos conformes a la ASC 842. Implementar una solución basada en NLP permitirá manejar la variabilidad en la estructura de los contratos y extraer de manera precisa la información clave necesaria para el cumplimiento normativo.

El objetivo principal de esta solución es proporcionar a los profesionales de la contabilidad una herramienta automatizada que les permita subir un archivo PDF de un contrato de arrendamiento y obtener respuestas concisas sobre la información relevante necesaria para la generación de un memorando técnico conforme a la norma ASC 842. Este sistema automatizado utilizará modelos avanzados de lenguaje natural (LLMs) para identificar y extraer automáticamente los campos de datos específicos previamente definidos, tales como términos de arrendamiento, pagos, opciones de renovación y otros elementos clave.

Actualmente, existe una aplicación web denominada gaapRT que permite a los profesionales contables ingresar manualmente la información de los contratos de arrendamiento. Sin embargo, este proceso es laborioso y propenso a errores. La automatización de este proceso a través del uso de NLP y LLMs no solo facilitará el trabajo de los profesionales contables al reducir significativamente el tiempo y el esfuerzo necesarios, sino que también mejorará la precisión y consistencia en la extracción de datos.

En resumen, los objetivos específicos de la solución son:

- Desarrollar un sistema que permita la carga de archivos PDF de contratos de arrendamiento.
- Utilizar modelos de NLP y LLMs para identificar y extraer automáticamente la información clave de estos contratos.
- Integrar esta funcionalidad en la aplicación web existente gaapRT para automatizar el proceso de generación de memorandos técnicos conforme a la ASC 842.
- Aumentar la eficiencia y precisión en la extracción de datos relevantes, minimizando los errores humanos y estandarizando el proceso.
- Evaluar y determinar la combinación óptima de Modelos de Lenguaje Grande (LLMs) que ofrezca la mejor relación costo-beneficio.

Esta solución permitirá a los profesionales contables centrarse en tareas de mayor valor añadido, confiando en la capacidad del sistema para manejar la variabilidad en la estructura de los contratos y extraer la información crítica de manera precisa y conforme a las normativas vigentes.

Específicamente, de los contratos de arrendamiento obtenidos, se tratará de extraer la siguiente información:

1. El nombre del arrendador (Pregunta 1)
2. La fecha de inicio del contrato (Pregunta 2)
3. La fecha de terminación del contrato (Pregunta 3)
4. La fecha efectiva del contrato (es decir, la fecha en la cual el contrato entra en vigor, muchas veces esta fecha coincide con la fecha de inicio del contrato) (Pregunta 4)
5. Si el arrendatario tiene la opción de comprar el objeto del arrendamiento (Pregunta 5)
6. Si el arrendatario tiene la opción de renovar el contrato o extenderlo más allá de la fecha de terminación del contrato (Pregunta 6)

7. Si el arrendatario tiene la opción de terminar el contrato de arrendamiento antes de la fecha de terminación del contrato (Pregunta 7)
8. Si el arrendador tiene la opción de terminar el contrato de arrendamiento antes de la fecha de terminación del contrato (Pregunta 8)

Para cada una de las respuestas, se busca una exactitud de al menos el 70%.

Investigación de datos

En esta sección del proyecto, se ha implementado un proceso automatizado para extraer datos relevantes sobre contratos de arrendamiento (lease agreements en inglés) que cumplen con la normativa ASC 842, utilizando la API de la SEC, los cuáles son de público acceso. Este método permite obtener información financiera contenida en los formularios 10-K, que son reportes anuales presentados por las empresas públicas estadounidenses. El objetivo es analizar los contratos de arrendamiento excluyendo enmiendas, adendos o extensiones. A continuación, se detallan los pasos realizados para esta extracción de datos.

Configuración de la consulta para obtener contratos del SEC

En este paso, se propone buscar términos relacionados con "lease agreements", excluyendo enmiendas, adendos y extensiones. La razón por la cual se hace esto es gaapRT como plataforma solo puede generar memorandos actualmente para contratos de arrendamientos nuevos ya que no se ha implementado la lógica para manejar los casos de enmiendas, adendos y extensiones.

La búsqueda está limitada a los formularios 10-K presentados entre el 1 de enero de 2020 y el 31 de diciembre de 2023.

Los parámetros de búsqueda definidos son los siguientes:

- **query**: "lease agreement between -amendment -addendum -extension", para obtener únicamente contratos de arrendamiento iniciales, evitando aquellos modificados o extendidos.
- **formTypes**: ["10-K"], para buscar únicamente formularios 10-K.
- **startDate** y **endDate**: Definen el período de búsqueda entre el 2020 y 2023.
- **page**: Define el parámetro de paginación de la API; la mayoría de los contratos se encontrarán en la primera página.

Una vez definida la consulta, se hace un llamado a la API para obtener las presentaciones que cumplen con los criterios establecidos. Para asegurar que los resultados sean relevantes, se filtró la data para eliminar aquellos registros que no contengan una descripción detallada del contrato de arrendamiento. Específicamente, se descartan filas cuya descripción es simplemente "10-K" o "FORM 10-K". Finalmente, se establece una opción que permite mostrar todos los resultados sin truncarlos, y se generó una tabla HTML interactiva que facilita la visualización y exploración de los datos. Esta visualización permite hacer clic en los enlaces de las presentaciones encontradas para inspeccionarlas más a fondo. A continuación, se muestra un ejemplo del resultado:

	companyNameLong	description	filingUrl
0	NeuBase Therapeutics, Inc. (NBSE) (CIK 0001173281)	EXHIBIT 10.29	https://www.sec.gov/Archives/edgar/data/1173281/000110465920003043/tm1927377d1_ex10-29.htm
1	Target Group Inc. (CBDY) (CIK 0001586554)	EXHIBIT 10.20	https://www.sec.gov/Archives/edgar/data/1586554/000110465920046312/tm205296d1_ex10-20.htm
2	JPMCC Commercial Mortgage Securities Trust 2017-JP7 (CIK 0001709967)	None	https://www.sec.gov/Archives/edgar/data/1709967/000188852423002975/jpc17jp7_10k-2022.htm
3	DCP Midstream, LP (DCP, DCP-PB, DCP-PC) (CIK 0001338065)	EX-99.2	https://www.sec.gov/Archives/edgar/data/1338065/000133806520000013/dcp-20191231exhibit992.htm
4	GLADSTONE INVESTMENT CORPORATION\DE (GAIN, GAINL, GAINM) (CIK 0001321741)	EX-99.3	https://www.sec.gov/Archives/edgar/data/1321741/000119312520140388/d914678dex993.htm
5	MARRIOTT INTERNATIONAL INC /MD/ (MAR) (CIK 0001048286)	EX-10.16	https://www.sec.gov/Archives/edgar/data/1048286/000162828023003485/mar-q42022xexx1016.htm

Figura 3 Ejemplo de metadata extraída con la API del SEC

Una vez que se obtuvo estos resultados, se procedió a analizar cada documento para verificar de forma manual que sean contratos de arrendamiento. Luego, se procedió a extraer un total de 100 contratos, convirtiéndolos de un formato HTML a un formato de PDF.

Luego de esto, se procede a obtener información sobre los PDFs extraídos. Una vez recolectado los contratos, se procedió a leerlos, analizarlos y responder a las preguntas planteadas anteriormente. Estas respuestas se pueden encontrar en la Tabla 10 de los anexos.

Las estadísticas del número de páginas y del número de palabras son las siguientes

Tabla 1 Estadísticas del número de páginas y número de palabras

	Número de páginas	Número de palabras
Media	22.88	10534.19
Desviación estándar	20.81	9668.32
Mínimo	2	779
Máximo	124	47873
Mediana	16	6976

Luego de tener una idea de la cantidad de páginas y palabras de los documentos, se procedió a estimar la cantidad de tokens de cada documento, para esto se procedió a utilizar la librería de Python llamada **tiktoken**, con la que se puede evaluar el número de tokens generados por los siguientes modelos de embeddings:

- text-embedding-ada-002
- text-embedding-3-small
- text-embedding-3-large

Estos 3 modelos de embeddings utilizan el tokenizador denominado **cl100k_base**.

Con esto, se obtienen las siguientes estadísticas del número de tokens:

Tabla 2 Estadísticas del número de tokens

	Número de tokens
Media	21157.36
Desviación estándar	19478.29
Mínimo	1602
Máximo	98598
Mediana	14070

A continuación, el histograma de los tokens por documento:

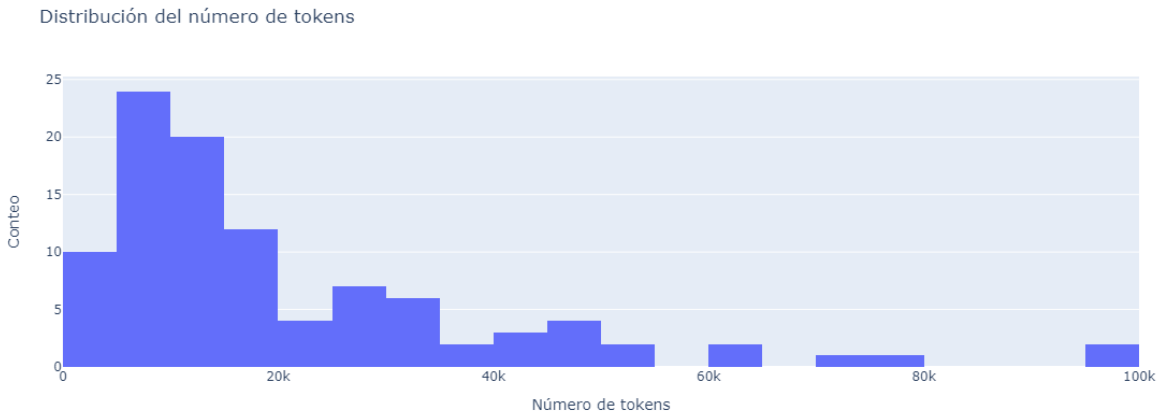

Figura 4 Histograma del número de tokens de los contratos obtenidos

El precio de cada uno de los 3 modelos de embedding tiene los siguientes precios por cada millón de tokens (en dólares):

- text-embedding-ada-002: $0.100 / 1M tokens
- text-embedding-3-large: $0.130 / 1M tokens
- text-embedding-3-small: $0.020 / 1M tokens

A partir de estos datos, se concluye que los precios (en dólares) para transformar los textos de los 100 PDFs en embeddings son los siguientes:

- Precio total para los 100 documentos usando text-embedding-ada-002: $0.21
- Precio total para los 100 documentos usando text-embedding-3-small: $0.04
- Precio total para los 100 documentos usando text-embedding-3-large: $0.28

También tenemos la lista de precios por utilizar los siguientes modelos:

- gpt-4o-mini: $0.150 / 1M input tokens
- gpt-4o-2024-08-06: $2.50 / 1M input tokens
- gpt-3.5-turbo: $0.500 / 1M input tokens

Con esta lista de precios y con la cantidad total de tokens que tenemos que utilizar los diferentes modelos de LLM para los 100 documentos para contestar las 8 preguntas cuesta lo siguiente:

- gpt-4o-mini: $2.56
- gpt-4o-2024-08-06: $42.32
- gpt-3.5-turbo: $8.48

Variabilidad de los contratos

A continuación, se muestra una muestra de cómo los contratos pueden presentar la respuesta a las preguntas dadas en diferentes formas.

Tabla 3 Muestra de la variabilidad de los contratos

Contrato	Texto
1	This Lease Agreement (“Lease”) is made and entered into effective as of this 25 day of January, 2022 (“Effective Date”) by and between BarBell Real Estate, LLC, a Texas limited liability company (“Landlord”) and HVE, Inc, a DelawareCorporation (“Tenant”) The term of this Lease Agreement (“Term”) shall commence on the Effective Date and expire on the five (5)year anniversary of the Rent Commencement Date. Tenant shall notify Landlord in writing of the expected Rent Commencement Date at least thirty (30) days prior to the Rent Commencement Date. If Landlord should be in default, Tenant shall have the option to terminate this Lease Agreement and be held harmless against any of its terms or obligations
2	Parties. GIFFORD INVESTMENTS, INC., a Massachusetts corporation, having an address at 111 South Worcester Street, Norton, Massachusetts 02712 (the “Lessor”) does hereby lease to CERAMICS PROCESS SYSTEMS CORPORATION, a Massachusetts corporation, having an address at 111 South Worcester Street, Norton, Massachusetts 02712 (the “Lessee”), and Lessee hereby leases the premises described in Section 2. Term. The term of this Lease shall be for ten (10) years, commencing as of March 1, 2006 and ending on February 29, 2016 (the “Term”). Lessee’s Option to Purchase. Lessee shall have the option to purchase the Leased Premises and the Property (the “Option”) from Lessor at anytime during the Term of the Lease. The purchase price shall be the Fair Market Value of the Leased Premises and Property, as hereinafter defined, but not less than $1,100,000, and shall be paid by Lessee to Lessor at the time of closing. If Lessee exercises the Option, Lessor and Lessee shall attempt to agree upon the Fair Market Value using their best good-faith efforts. If Lessor and Lessee fail to reach an agreement within thirty (30) days following Lessee’s exercise of the Option (the “Outside Agreement Date”), then each party shall make a separate determination of the Fair Market Value which shall be submitted to each other and to arbitration in accordance with the following items (i) through (vii)

3	THIS LEASE AGREEMENT (this “Agreement”) is made and entered into this September 24, 2021 (the “Effective Date”), by and between CARGILL, INCORPORATED, a Delaware corporation, as landlord (“Landlord”), and SUSTAINABLE OILS, INC., a Delaware corporation, as tenant (“Tenant”) Term of the Lease. The term of the Lease shall be for a period of sixty (60) months, commencing on November 1, 2021 (the “Commencement Date”) and expiring on October 31, 2026 (the “Lease Term” or “term of the Lease”). Each twelve-month period commencing on the Commencement Date and each one-year period thereafter, is hereinafter referred to as a “Lease Year”. Upon the written request by Tenant, the Lease may be terminated prior to the end of the Lease Term in order to facilitate Tenant’s purchase of the Premises as set forth in Section 16. No right of renewal is expressed or implied in the Lease. The “Effective Time” of the Commencement Date will be 12:01 a.m. Mountain time on the Commencement Date for all purposes of the Lease.

En todos los casos, se evidencia que la información que se quiere extraer viene escrita de distintas formas ya que no existe un estándar para escribir estos contratos.

Preparación de los datos

Luego de estimar los precios de generar embeddings para los distintos modelos de embedding, se procede a guardar cada uno de los 100 documentos en un almacenamiento en la nube de AWS, específicamente en AWS S3

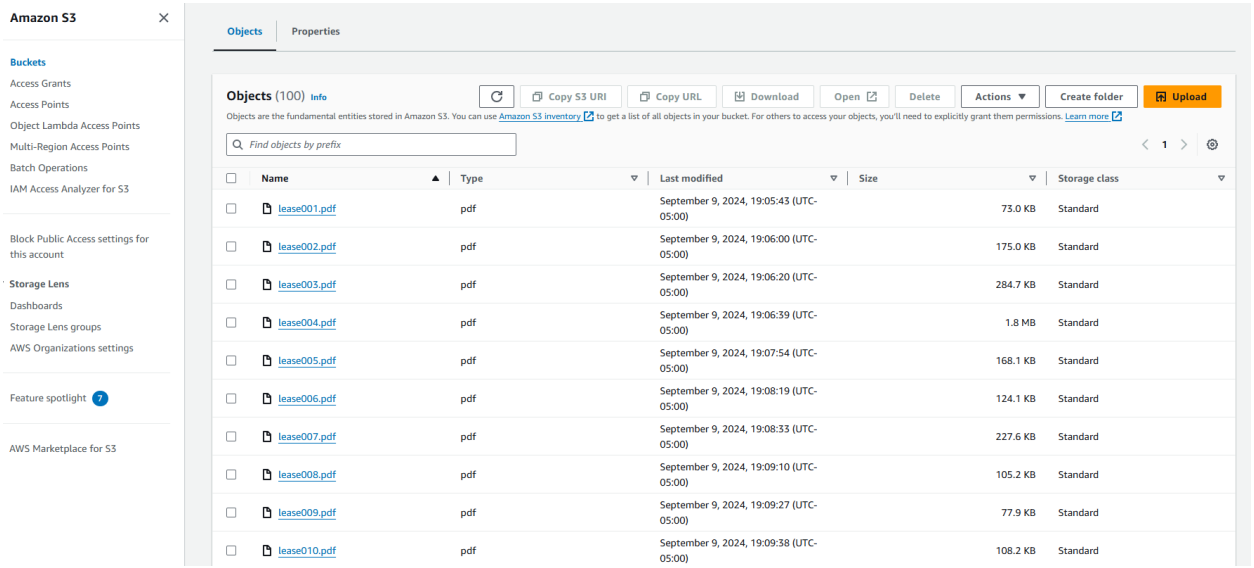

Figura 5 Documentos guardados en AWS S3

El siguiente paso es cargar estos documentos de S3, dividir el texto utilizando una técnica recursiva.

Específicamente, el **RecursiveCharacterTextSplitter** de Langchain es una herramienta recomendada para dividir texto de manera efectiva, especialmente en aplicaciones que requieren un procesamiento de texto genérico. Este splitter se parametriza mediante una lista de caracteres que se utilizan como delimitadores para realizar las divisiones en el texto.

El splitter intenta dividir el texto utilizando los caracteres de la lista en orden, hasta que los fragmentos resultantes alcanzan un tamaño adecuado. La lista de delimitadores por defecto incluye:

- "\n\n": intenta mantener la integridad de los párrafos.
- "\n": busca dividir el texto en líneas.
- " ": divide el texto en palabras.
- "": permite una división adicional a nivel de caracteres.

Este enfoque tiene como objetivo conservar juntos los fragmentos de texto que están semánticamente relacionados, como párrafos, oraciones y palabras, hasta que se alcanza el tamaño deseado para los chunks. De esta manera, se maximiza la coherencia y la relación contextual de los datos que se procesan.

El **RecursiveCharacterTextSplitter** es especialmente útil en contextos donde se necesita una representación precisa y coherente del contenido textual, como en tareas de análisis de texto, generación de resúmenes y sistemas de recuperación de información.

Una vez que el texto fue dividido, se guarda cada split como un vector en una base de datos. En este caso, será Pinecone.

Es muy importante que las dimensiones necesarias para text-embedding-ada-002 y para text-embedding-3-small es de **1536** mientras que para text-embedding-3-large es de **3072**, es decir el doble. De esta forma, se crean dos índices con estas dimensiones

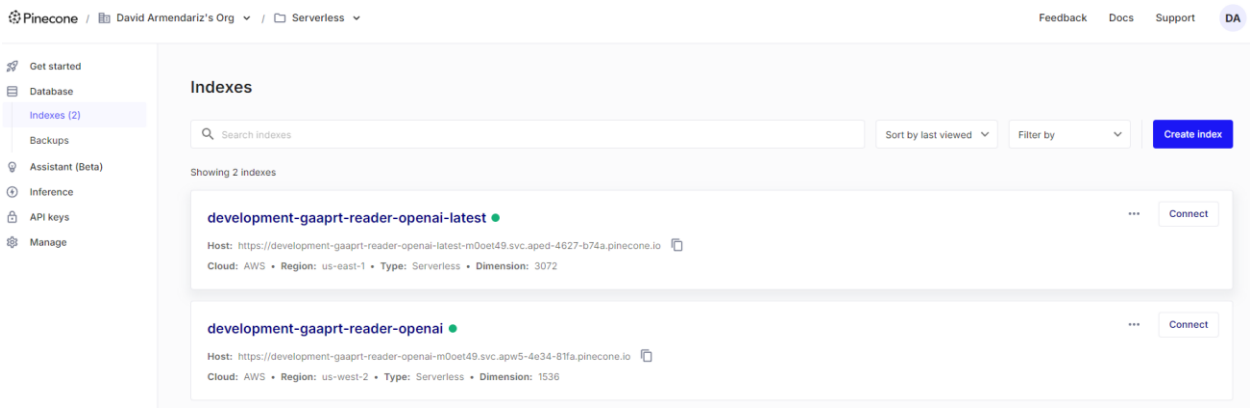

Figura 6 Índices creados para Pinecone

Cada documento se guarda en lo que Pinecone denomina un **namespace**, para poder agruparlos por documento

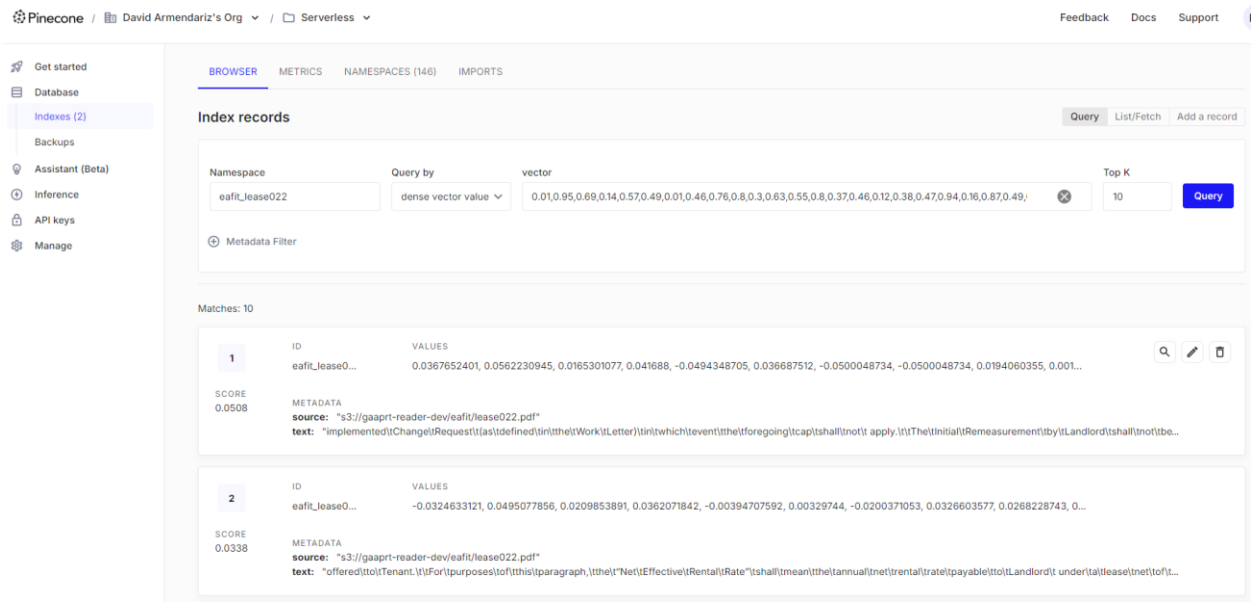

Figura 7 Ejemplo de vectores guardados en namespaces

Desarrollo

Todo el código fuente del desarrollo se puede encontrar en el siguiente repositorio público: https://github.com/DavidArmendariz/eafit-tesis

El artefacto que se desarrollará consiste en un sistema integral que cumplirá con los siguientes componentes y funcionalidades:

Interfaz de usuario

Se desarrolló una interfaz gráfica en React que permite a los usuarios subir archivos PDF desde sus computadoras. La interfaz también proporciona feedback al usuario sobre el estado de la carga.

A continuación, se muestra el componente gráfico que permite adjuntar un documento:

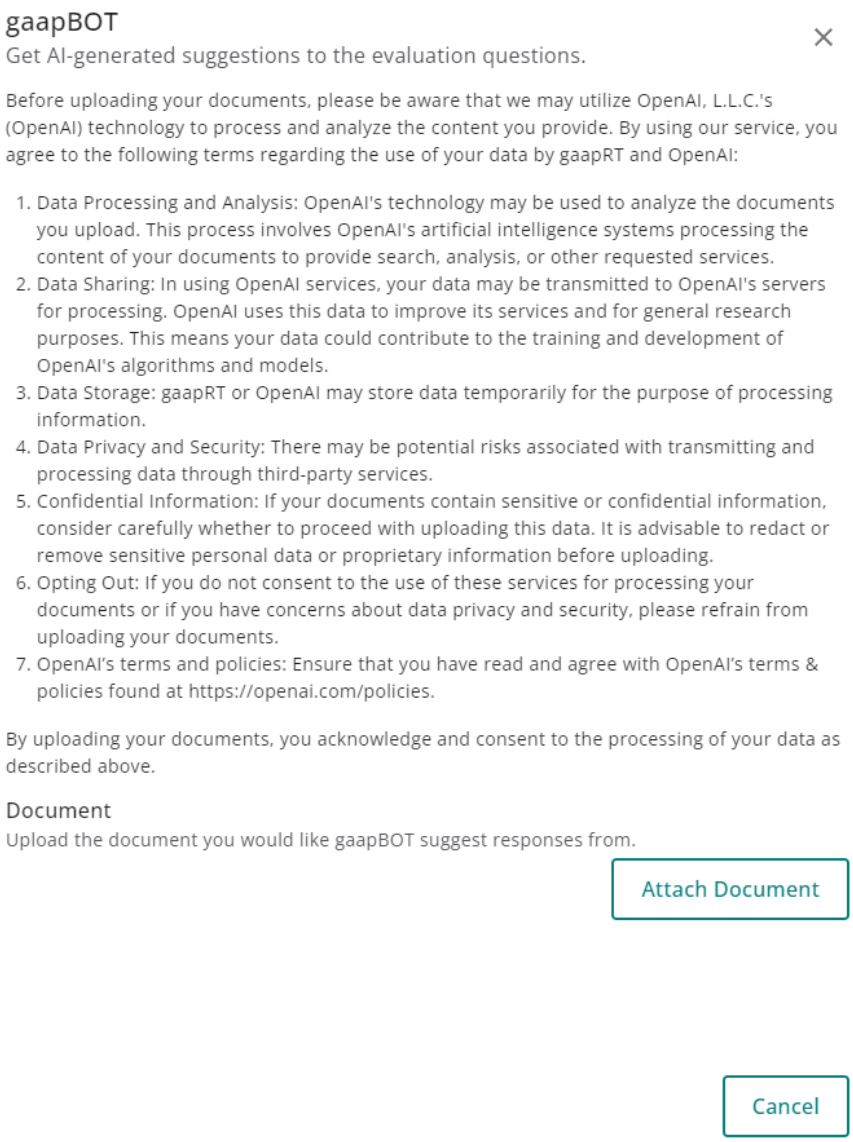

Figura 8 UI para adjuntar un documento

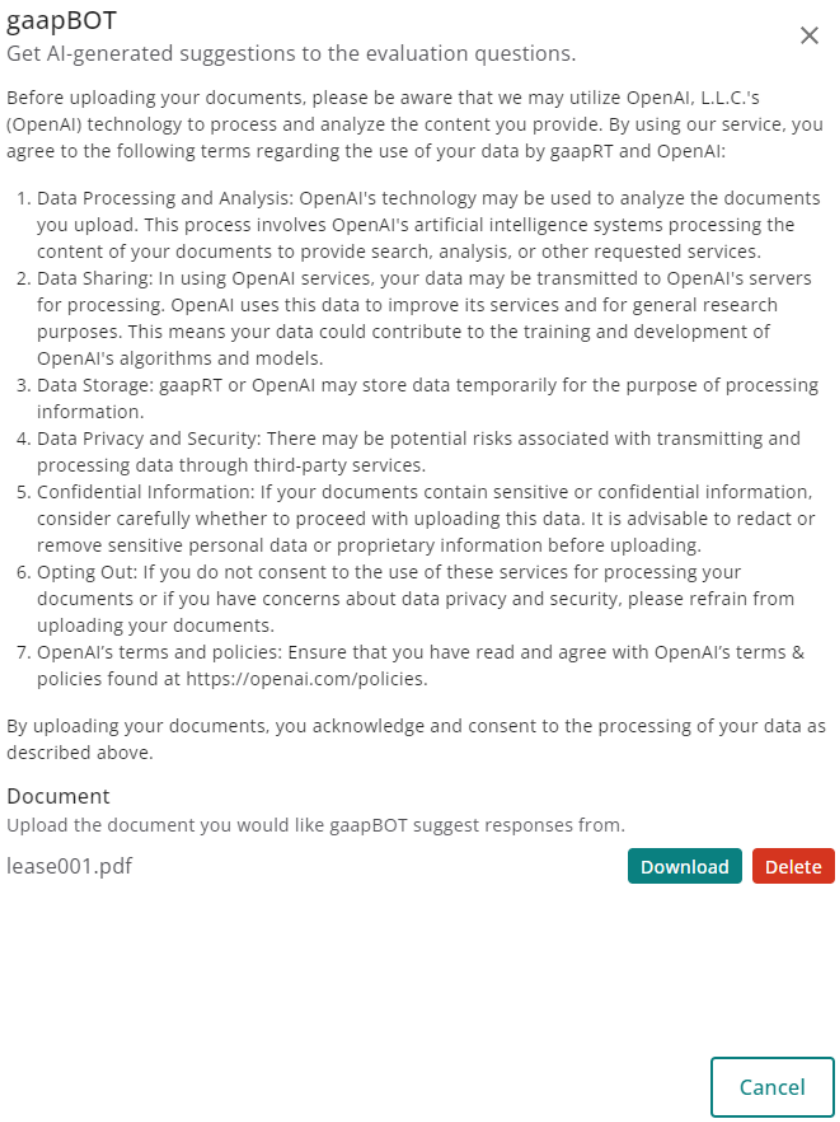

Figura 9 UI para cuando el documento ya ha sido subido

API de recepción y almacenamiento

Se implementó una API que recibe el archivo PDF cargado por el usuario y lo almacena en un sistema de almacenamiento en la nube (Amazon S3). Esto se logró con el framework Ruby on Rails. Se recibe el PDF como un blob y se lo guarda en un bucket de S3, asegurando que la metadata esté guardado en una base de datos para que sea accesible para su procesamiento posterior.

API de procesamiento de archivos

Se desarrolló una API que lee el archivo almacenado en S3, lo procesa en fragmentos manejables (chunks) utilizando la técnica de **RecursiveCharacterTextSplitter** descrita anteriormente y utiliza un modelo "encoder" para convertir estos chunks en embeddings. Por último, almacena los embeddings generados en Pinecone. Los modelos “encoder” para generar los embeddings fueron los siguientes:

- text-embedding-ada-002
- text-embedding-3-large
- text-embedding-3-small

API de generación de respuestas

Se desarrolló una API que lee los embeddings almacenados en la base de datos vectorial y los pasa a un modelo "decoder" para generar respuestas específicas en el contexto de la norma ASC 842. Los modelos “decoder” utilizados fueron los siguientes:

- gpt-3.5-turbo
- gpt-4o-mini
- gpt-4o-2024-08-06

Experimentos y prompts utilizados

Se hizo varios experimentos utilizando diferentes combinaciones. Inicialmente, no se utilizó salidas estructuradas debido a que esta funcionalidad no estaba disponible al público.

A continuación, se describen los experimentos. Los experimentos incrementan en complejidad.

Tabla 4 Experimentos realizados

Número de experimento	Modelo LLM	Modelo de embedding	¿Utiliza salidas estructuradas?
1	gpt-3.5-turbo	text-embedding-ada-002	No
2	gpt-3.5-turbo	text-embedding-3-small	No
3	gpt-3.5-turbo	text-embedding-3-large	No
4	gpt-3.5-turbo	text-embedding-ada-002	Sí
5	gpt-3.5-turbo	text-embedding-3-small	Sí
6	gpt-3.5-turbo	text-embedding-3-large	Sí
7	gpt-4o-mini	text-embedding-3-small	No
8	gpt-4o-mini	text-embedding-3-large	No
9	gpt-4o-mini	text-embedding-3-small	Sí
10	gpt-4o-mini	text-embedding-3-large	Sí
11	gpt-4o-2024-08-06	text-embedding-3-small	Sí
12	gpt-4o-2024-08-06	text-embedding-3-small	Sí

El uso de salidas estructuradas en modelos de lenguaje como **gpt-4o-mini** y **gpt-4o-2024-08-06** es fundamental para maximizar su efectividad en aplicaciones prácticas. Este enfoque garantiza que las respuestas sigan un esquema predefinido, lo que facilita la comprensión y el análisis de la información. Además, permite una integración sencilla en sistemas automatizados, ya que formatos como JSON o XML son fácilmente procesables por otras aplicaciones. Al adaptar las respuestas a formatos específicos, como listas o tablas, se mejora la experiencia del usuario y su satisfacción. Asimismo, las salidas estructuradas facilitan ajustes a diferentes contextos y necesidades, optimizando la relevancia de la información proporcionada. Por último, un formato claro simplifica el proceso de revisión y post-edición, ahorrando tiempo y recursos. En resumen, las salidas estructuradas mejoran la claridad, la utilidad y la integración de las respuestas generadas, convirtiéndolos en una práctica recomendada en el uso de modelos de lenguaje.

Por estas razones se utilizarán solamente outputs estructurados en los experimentos que involucran **gpt-4o-mini** y **gpt-4o-2024-08-06**.

En este proyecto, se utilizó un prompt que no usó salidas estructuradas y un prompt que sí las usó. El prompt que no hizo uso de esta funcionalidad es el siguiente:

Your task is to read the provided Source Document and respond to the questions provided in Inquiries. Inquiries is an array of objects, each representing a question and its parameters. Each object in the Inquiries array contains the following keys:

reader_question: the value notes the question you have to answer;

id: the value is the question's id;

restrictions: the value notes any requirements for or restrictions on the response you can provide for that specific question.

user_response: the value is optional. The reader_question will explain how this value is to be utilized.

type: the value represents the expected format of the <<answer>> you provide.

Expected format of the <<answer>>:

Your response must be provided as a JSON object in the following format - each question should be identified by its question id:

{{"<<the question's id>>": {{"question": "<<reader_question>>", "answer": "<<answer>>", "meta": "<<meta>>"}} }}

- If type='string', then provide the <<answer>> in the form of an array of strings in which each paragraph is a separate string in the array. If the question asks to provide the response in the form of

a list, then the first string in the array should be an introduction of the list and the remaining strings should be the list.

- If type='boolean', then the <<answer>> can only be "0" or "1". Reply with "0" if false, "1" if true, or nothing if unknown or there is an error.

- If type='number', then only provide the <<answer>> as a number or nothing if unknown or there is an error.

- If type='date', then only provide the <<answer>> as date formatted as YYYY-MM-DD or nothing if unknown or there is an error.

If required in the response, refer to the Source Document as the "source document."

In <<meta>>, where possible, you should first quote the sentence where you found this answer.

And then, if possible, you should explain how you determined your answer. But this is not required. If nothing is provided in <<meta>>, then leave it blank.

If your response does not meet the following requirements then fix it:

- Your answer must come from the provided Source Document. Remove any information from outside the Source Document.

- You must respond to all questions in the Inquiries array.

- Return nothing in the <<answer>> if there is any type of error, you have no response or you do not know the answer to the question.

- Confirm that all the restrictions associated with a question are enforced.

- Do not mention, disclose or reference the restrictions associated with a question anywhere in the response you provide.

Inquiries: {questions}

Source Document: {context}

Figura 10 Prompt sin outputs estructurados

El prompt que sí utilizó salidas estructuradas es el siguiente:

Your task is to read the provided Source Document and respond to the questions provided in Inquiries.

Inquiries is an array of objects, each representing a question and its parameters.

Each object in the Inquiries array contains the following keys:

- id: the value is the question's id;

- reader_question: the value notes the question you have to answer;

- restrictions: the value notes any requirements for or restrictions on the response you can provide for that specific question.

- user_response: the value is optional. The reader_question will explain how this value is to be utilized.

- type: the value represents the expected format of the <<answer>> you provide.

- If type='string', then provide the <<answer>> in the "answer_string" key in the form of an array of strings in which each paragraph is a separate string in the array. If the question asks to provide the response in the form of a list, then the first string in the array should be an introduction of the list and the remaining strings should be the list.

- If type='boolean', then the <<answer>> in the "answer_boolean" key. It can only be "0" or "1". Reply with "0" if false, "1" if true.

- If type='date', then provide the <<answer>> in the "answer_date" key as a date formatted as YYYY-MM-DD.

In <<meta>>, where possible, you should first quote the sentence where you found this answer. And, if possible, you should explain how you determined your answer. If your response does not meet the following requirements then fix it:

- Your answer must come from the provided Source Document. Remove any information from outside the Source Document.

- You must respond to all questions in the Inquiries array.

- Return nothing in the <<answer>> if there is any type of error, you have no response or you do not know the answer to the question.

- Confirm that all the restrictions associated with a question are enforced.

- Do not mention, disclose or reference the restrictions associated with a question anywhere in the response you provide.

Inquiries: {questions}

Source Document: {context}

Figura 11 Prompt con outputs estructurados

En ambos casos, el documento entero es inyectado como contexto. También, para cada pregunta, se inyecta el objecto “inquiries”. Esto no es más que la pregunta que se va a hacer.

Las preguntas o “inquiries” utilizadas en cada caso son las siguientes:

Tabla 5 Preguntas utilizadas para cada pregunta

Pregunta	**Pregunta para el prompt**
1	What is the name of the lessor in the agreement in the source document? The lessor is also referred to as the "landlord" in certain cases. Return only the exact name of the lessor. Format any all-caps values in the text to title case. Write the response without using definitional terms. Refer to each entity or concept by its full name or a clear descriptor each time it is mentioned, rather than assigning it a shorthand label or nickname.
2	On what date was the contract executed? This could the date the document was signed, or if that date is unavailable, it’s the date on which the agreement was entered into or is dated.
3	What is the lease commencement date? The commencement date is the date on which a lessor makes an underlying asset available for use by a lessee.
4	According to the contract, when is the last day of the lease (i.e., the date on which the lessee will lose the right to use the assets)?
5	Does the lease contract grant the lessee the option to purchase any assets?
6	Does the lease contract grant the lessee the option to extend or renew the lease beyond it stated term?
7	Does the lease contract grant the lessee the option to terminate or cancel the lease before it stated term?
8	Does the lessor have the right to terminate or cancel the lease prior to the expiration date?

Evaluación

La evaluación del sistema se centrará en medir la precisión, la completitud y la conformidad de los datos extraídos. Se utilizará la técnica de **exact match** como se describió en el marco teórico. La métrica que se utilizará será el **accuracy**, para las preguntas 2, 3, 4, 5, 6, 7 y 8.

Para la pregunta 1, se utilizará una métrica diferente, utilizando la **distancia de Levenshtein** con un umbral del 80%. La justificación del por qué se da más flexibilidad a esta pregunta, es que el nombre del arrendador puede componerse muchas veces del nombre, junto con algún acrónimo como “LLC”. Que incluya o no el acrónimo no es muy relevante.

Plan de evaluación

Se utilizará la muestra representativa de contratos de arrendamiento en PDF, abarcando diferentes formatos y estructuras para asegurar la generalización del sistema. En este caso, se seleccionaron 100 contratos.

Para la iteración y mejora, primero se evaluarán los 100 contratos utilizando la técnica de zero-shot prompting. Una vez hecho esto, se calcularán las métricas y se reportarán en formatos tabulares. Se escogió esta técnica debido a la variabilidad de los contratos y su necesidad de categorizarlos según su complejidad. La primera opción que se evaluó fue de contratar a gente experta en el tema y que los clasificara según su nivel de complejidad. Sin embargo, hubo muchas limitaciones tanto económicas como de tiempo para que este enfoque sea viable. Aparte, la complejidad percibida por un ser humano es relativa. Por eso se delegó a la inteligencia artificial calcular el nivel de complejidad de los contratos para poder estratificar los contratos del conjunto de prueba.

Luego, para garantizar una evaluación robusta y representativa, se implementará una estrategia de particionamiento estratificado por la complejidad de los contratos, según qué tan difícil es encontrar la respuesta a las preguntas planteadas. El conjunto de datos original se dividirá en dos subconjuntos: uno de validación y otro de prueba, manteniendo una proporción del 70% y 30% respectivamente. El conjunto de pruebas no será utilizado para la optimización de prompts. Solo para la evaluación final de métricas.

Se procedió con este enfoque debido a las limitaciones en la obtención de más contratos a través de la API de SEC, ya que el acceso a una mayor cantidad de documentos resulta complicado por restricciones de disponibilidad y tiempo. Además, el proceso de lectura y extracción de información relevante de estos contratos es altamente demandante en términos de tiempo, lo cual dificulta considerablemente la expansión del conjunto de datos en un plazo razonable. Por ello, la selección de 100 contratos como muestra representativa, junto con el uso de técnicas como el zero-shot prompting y la partición estratificada, asegura un balance adecuado entre viabilidad y validez científica en la evaluación del sistema.

La estratificación se realizará utilizando una variable derivada que categoriza los registros según el conteo de aciertos por documento en la primera iteración donde se evalúan los 100 contratos. Esta aproximación asegura que la distribución de las categorías de precisión se mantenga consistente tanto en el conjunto de entrenamiento como en el de prueba. Al preservar la representatividad de las categorías, se mitigará el riesgo de sesgos en la evaluación del modelo.

Se establecerá un estado aleatorio fijo para garantizar la reproducibilidad del proceso de división. Esto permitirá que otros investigadores puedan replicar exactamente la misma partición de datos, lo cual es crucial para la validez científica del estudio.

Los conjuntos de datos resultantes serán almacenados en archivos separados: un conjunto completo con todos los registros, y dos subconjuntos específicos para validación y prueba.

Procedimientos de Evaluación

Testeo automatizado

Se implementará un script para evaluar automáticamente todos los contratos y calcular las métricas descritas anteriormente.

Análisis de errores

Se realizará un análisis de los errores identificados durante la evaluación, categorizándolos y determinando las causas raíz para iterar y mejorar los resultados.

Reporte de resultados

Se documentarán todos los hallazgos de la evaluación. Específicamente se reportarán: las métricas de rendimiento para los 100 contratos iniciales, la descripción de las mejoras para cada pregunta y las métricas de estas mejoras en el conjunto de validación y de test.

Resultados

A continuación, se muestran los resultados para cada experimento y por cada pregunta. Cada experimento consiste en evaluar los 100 contratos y calcular el accuracy.

Tabla 6 Resultados de los distintos experimentos

	Experimento											
Pregunta	1	2	3	4	5	6	7	8	9	10	11	12
1	**0,5800**	0,4000	0,4300	0,5700	0,4100	0,4400	0,4200	0,4800	0,4300	0,4700	0,4300	0,4700
2	0,4362	0,3191	0,2447	0,4149	0,3404	0,2340	0,5000	**0,5213**	**0,5213**	0,3511	**0,5213**	0,3723
3	0,5955	0,6291	0,5393	0,5730	0,6291	0,5169	0,6404	0,6180	**0,6629**	**0,6629**	0,6404	0,6517
4	0,3181	0,3750	0,4091	0,3750	0,3863	0,4091	0,4318	0,5114	0,4659	**0,5341**	0,4318	0,4545
5	0,9192	0,8990	0,9394	0,9293	0,8788	0,8990	0,9394	0,9394	0,9495	**0,9596**	0,9495	**0,9596**
6	0,8061	0,8469	**0,8673**	0,8265	0,8571	0,8367	0,8571	0,8469	0,8571	0,8367	0,8571	0,8163
7	0,5408	0,6224	0,6735	0,5408	0,6224	**0,6837**	0,5510	0,6429	0,5306	0,5918	0,6735	0,6429
8	0,6633	0,6837	0,8265	0,6429	0,6531	0,8163	0,6020	0,8470	0,5918	0,8776	0,7653	**0,9388**

Luego, se procedió a calcular cuál fue el costo por pregunta y por experimento utilizando las estimaciones calculadas anteriormente. Los valores de cada celda están en dólares americanos. Es decir, cada valor es el costo de haber evaluado los 100 contratos en su conjunto.

Tabla 7 Costo por experimento y por pregunta en dólares

	Experimento											
Pregunta	1	2	3	4	5	6	7	8	9	10	11	12
1	1.06	1.06	1.06	1.06	1.06	1.06	0.32	0.32	0.32	0.32	5.29	5.29
2	1.06	1.06	1.06	1.06	1.06	1.06	0.32	0.32	0.32	0.32	5.29	5.29

3	1.06	1.06	1.06	1.06	1.06	1.06	0.32	0.32	0.32	0.32	5.29	5.29
4	1.06	1.06	1.06	1.06	1.06	1.06	0.32	0.32	0.32	0.32	5.29	5.29
5	1.06	1.06	1.06	1.06	1.06	1.06	0.32	0.32	0.32	0.32	5.29	5.29
6	1.06	1.06	1.06	1.06	1.06	1.06	0.32	0.32	0.32	0.32	5.29	5.29
7	1.06	1.06	1.06	1.06	1.06	1.06	0.32	0.32	0.32	0.32	5.29	5.29
8	1.06	1.06	1.06	1.06	1.06	1.06	0.32	0.32	0.32	0.32	5.29	5.29

A partir de esto, se hace un análisis de costo beneficio, siendo el beneficio la métrica de exactitud o accuracy y siendo el costo el costo por experimento y por pregunta en dólares. Es decir, dividimos cada entrada de la

Tabla 6 con cada entrada de la Tabla 7. Entre más alto sea esta métrica, es mejor. Esto da como resultado la siguiente matriz de costo beneficio. En la última fila se muestran los dos mejor puntuados.

Tabla 8 Matriz de costo beneficio

	Experimento											
Pregunta	1	2	3	4	5	6	7	8	9	10	11	12
1	0.54717	0.37736	0.40566	0.53774	0.38679	0.41509	1.3125	1.5	1.34375	1.46875	0.08129	0.08885
2	0.41151	0.30104	0.23085	0.39142	0.32113	0.22075	1.5625	1.62906	1.62906	1.09719	0.09854	0.07038
3	0.56179	0.59349	0.50877	0.54057	0.59349	0.48764	2.00125	1.93125	2.07156	2.07156	0.12106	0.12319
4	0.30009	0.35377	0.38594	0.35377	0.36443	0.38594	1.34938	1.59813	1.45594	1.66906	0.08163	0.08592
5	0.86717	0.84811	0.88623	0.8767	0.82906	0.84811	2.93563	2.93563	2.96719	2.99875	0.17949	0.1814
6	0.76047	0.79896	0.81821	0.77972	0.80858	0.78934	2.67844	2.64656	2.67844	2.61469	0.16202	0.15431
7	0.51019	0.58717	0.63538	0.51019	0.58717	0.645	1.72188	2.00906	1.65813	1.84938	0.12732	0.12153
8	0.62575	0.645	0.77972	0.60651	0.61613	0.77009	1.88125	2.64688	1.84938	2.7425	0.14467	0.17747
Promedio	0.57302	0.56311	0.58134	0.57458	0.56335	0.57025	1.93035	2.11207	1.95668	2.06398	0.1245	0.12538

Según estos resultados, el mejor experimento fue el Experimento 8. Sin embargo, este experimento no utiliza salidas estructuradas, lo cual es un problema para el desarrollo de cualquier aplicación que utilice inteligencia artificial. Dado a que la consistencia es extremadamente importante, se considera que el segundo mejor experimento, es decir, el Experimento 10, es con el cual se va a trabajar de ahora en adelante.

Se calcula la diferencia absoluta entre el mejor resultado de cada pregunta y el resultado del Experimento 10 para dicha pregunta.

Tabla 9 Diferencia absoluta entre resultados del Experimento 10 y el mejor de cada fila

Pregunta	Diferencia entre Experimento 10 y el mejor de cada fila
1	0.11
2	0.1702
3	0
4	0
5	0
6	0.0306
7	0.0919
8	0.0612

Con estos resultados, se procede a realizar varios experimentos adicionales para mejorar los resultados por cada pregunta. Pero primero, se hizo la partición de la data en un conjunto de validación y uno de test estratificando por el total de aciertos de preguntas por contrato. Se realizó el siguiente histograma:

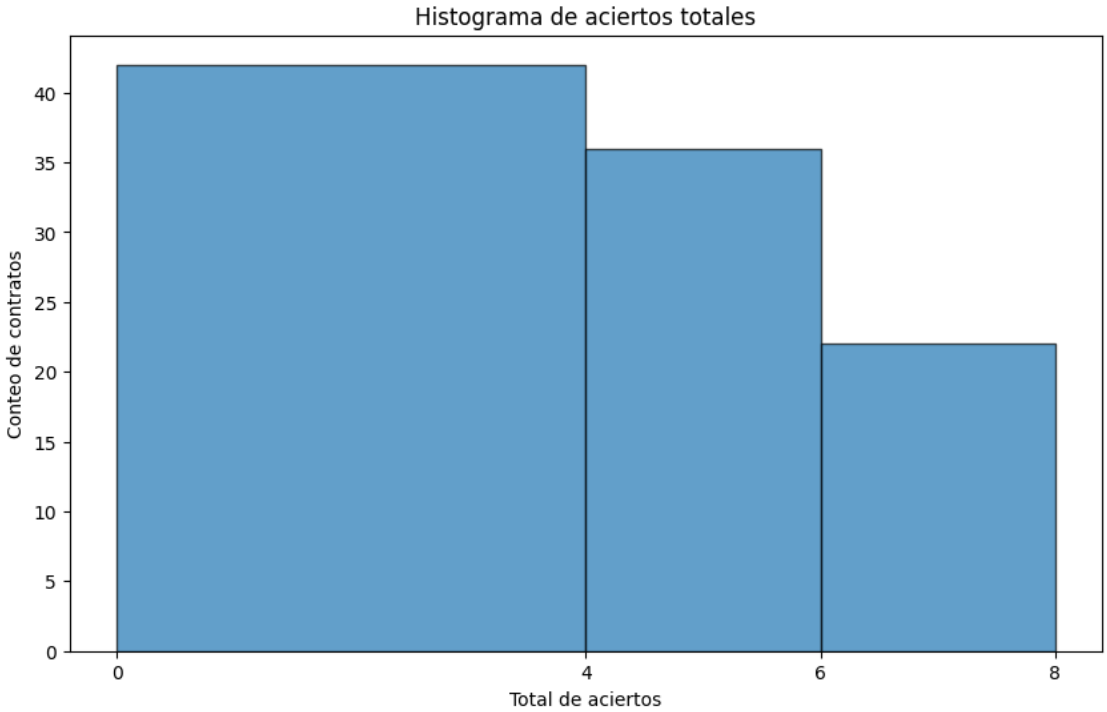

Figura 12 Histograma del total de aciertos de preguntas

Esta estratificación fue la mejor para tener una distribución relativamente igualitaria entre clases. Es decir, los contratos se clasificaron en las siguientes clases:

- Clase 1 (contratos difíciles): se acertaron entre 0 y 4 preguntas
- Clase 2 (contratos medios): se acertaron entre 5 y 6 preguntas
- Clase 3 (contratos fáciles): se acertaron entre 7 y 8 preguntas

Mejoras para la pregunta 1

Para la pregunta 1, se hicieron distintas variaciones para mejorar la pregunta. A continuación se muestra un resumen de las variantes y sus resultados:

Variación	Descripción	Validation accuracy	Test accuracy
1	Cambiar la pregunta para evitar ciertas respuestas	0.4857	0.4000
2	Cambiar la pregunta utilizando la técnica de Few Shot Prompting	0.4857	0.4667
3	Simplificar la pregunta de la variación 1	0.4143	0.3667
4	Cambiar el prompt y la pregunta	0.4286	0.4000
5	Utilizar LLM Chain Filters para la compresión contextual	0.4000	0.4000
6	Utilizar LLM Chain Filters usando la pregunta como prompt	0.5714	0.6333
7	Utilizar LLM Chain Filters haciendo referencia al esquema de la salida estructurada	0.5714	**0.6333**
8	Usar la variación 7 cambiando la temperatura a 0.1 del LLM	0.5857	0.6333
9	Utilizar LLM Listwise Rerank para la compresión contextual	0.6143	0.6000
10	Utilizar Embedding Filters	0.0000	0.0000

Variación 1: Cambiar la pregunta para evitar ciertas respuestas

En esta variación, la pregunta que se inyecta en el prompt se cambió por la siguiente:

> What is the name of the lessor in the agreement in the source document? The lessor is also referred to as the "landlord" in certain cases. Return only the exact name of the lessor. "Landlord" and "Lessor" are not valid answers. Format any all-caps values in the text to title case. Write the response without using definitional terms. Refer to each entity or concept by its full name or a clear descriptor each time it is mentioned, rather than assigning it a shorthand label or nickname.

En este prompt, se especificó que “landlord” y “lessor” no son respuestas válidas ya que en muchas ocasiones, estas eran las respuestas que arrojaba el LLM.

Variación 2: Cambiar la pregunta utilizando few shot prompting

El prompt utilizado para aplicar esta técnica al caso en particular de la pregunta 1 es el siguiente:

> What is the name of the lessor in the agreement in the source document? The lessor is also referred to as the "landlord" in certain cases. For example in this text I want you to give me "My LLC": "This

lease agreement (the “Lease”) made by and between MY LLC (the “Landlord”)". Return only the exact name of the lessor. "Landlord" and "Lessor" are not valid answers. Format any all-caps values in the text to title case. Write the response without using definitional terms. Refer to each entity or concept by its full name or a clear descriptor each time it is mentioned, rather than assigning it a shorthand label or nickname.

Variación 3: Simplificar la pregunta de la variación 1

Se modificó para que haga exactamente lo mismo que la variación 1, pero con menos palabras para que de esta forma el LLM se pueda enfocar. El nuevo prompt pasó a ser este:

What is the exact name of the lessor in the source document agreement? Note that the lessor may also be called the 'landlord'. Return only the full name of the lessor, formatted in title case if it appears in all caps. Exclude 'landlord' or 'lessor' from your response. Use full names for all entities mentioned.

Variación 4: Cambiar el prompt y la pregunta

La pregunta utilizada se inyecta en el prompt de la Figura 11. Debido a que este prompt se encarga de lidiar con distinto tipo de preguntas a la vez, se procedió a cambiar este prompt por algo que sea específico para la pregunta 1. El siguiente prompt se utilizó:

Your task is to read the provided Source Document and answer the questions provided in Inquiries. Inquiries is an object representing a question and its parameters. The relevant keys are the following:

- id: the question's id;

- reader_question: the question you have to answer;

- restrictions: any requirements for or restrictions on the response you can provide for that specific question.

Provide the answer in the "answer_string" key in the form of an array of strings in which each paragraph is a separate string in the array. If the question asks to provide the response in the form of a list, then the first string in the array should be an introduction of the list and the remaining strings should be the list. In "meta", where possible, you should first quote the sentence where you found this answer. And, if possible, you should explain how you determined your answer.

If your response does not meet the following requirements then fix it:

- Your answer must come from the provided Source Document. Remove any information from outside the Source Document.

- Return nothing if there is any type of error, you have no response or you do not know the answer to the question.

- Confirm that all the restrictions associated with a question are enforced.

- Do not mention, disclose or reference the restrictions associated with a question anywhere in the response you provide.

Inquiries: {questions}

Source Document: {context}

La pregunta utilizada fue la siguiente:

What is the exact name of the lessor in the source document agreement? Note that the lessor may also be called the 'landlord'. Return only the full name of the lessor, formatted in title case if it appears in all caps. Exclude 'landlord' or 'lessor' from your response. Use full names for all entities mentioned.

Variación 5: Utilizar LLM Chain Filters para achicar el tamaño del contexto

Un desafío con RAG es que, por lo general, no se conocen las consultas específicas que enfrentará el sistema de almacenamiento de documentos cuando se ingresa información en él. Esto significa que la información más relevante para una consulta podría estar oculta en un documento con mucho texto irrelevante. Pasar ese documento completo a través de la aplicación puede resultar en llamadas a modelos de lenguaje (LLM) más costosas y respuestas de menor calidad.

La compresión contextual está diseñada para solucionar esto. La idea es simple: en lugar de devolver los documentos recuperados tal como están, se puede comprimirlos utilizando el contexto de la consulta dada, de modo que solo se devuelva la información relevante. Comprimir se refiere tanto a reducir el contenido de un documento individual como a filtrar documentos completos que no sean útiles.

El LLMChainFilter es un compresor simple pero robusto. Utiliza un LLM para decidir cuáles de los documentos recuperados inicialmente filtrar y cuáles devolver, sin manipular el contenido de los documentos.

En esta variación 5, se utilizó esta técnica. A la pregunta para obtener el contexto de la base de datos vectorial se le denominará **pregunta de contexto** (retriever question en inglés). En este caso la pregunta de contexto fue la siguiente:

Get the information about the parties of the agreement

El prompt también fue cambiado a algo más sencillo. Este es el prompt que se utilizó ahora:

What is the name of the lessor? Provide only the name of the entity.

Context: {context}

Variación 6: Utilizar LLM Chain Filters usando la pregunta como prompt

Como se evidencia en la variación 5, el prompt se simplificó de manera sustancial, ya que el contexto también se simplificó. En esta variación se utilizó la misma pregunta de contexto pero se modificó el prompt. El prompt utilizado fue el siguiente:

What is the name of the lessor in the agreement in the Context? The lessor is also referred to as the "landlord" in certain cases. Return only the exact name of the lessor. Format any all-caps values in the text to title case. Write the response without using definitional terms. Refer to each entity or concept by its full name or a clear descriptor each time it is mentioned, rather than assigning it a shorthand label or nickname.

Context: {context}

El prompt ahora es básicamente la pregunta que se utilizó originalmente.

Variación 7: Utilizar LLM Chain Filters haciendo referencia al esquema de la salida estructurada

La variación 7 es básicamente la misma que la variación 6, con la diferencia de que en el prompt se hace referencia al esquema de la salida estructurada. El esquema contiene un tipo llamado “answer_string” que es del tipo string.

What is the name of the lessor in the agreement in the Context? The lessor is also referred to as the "landlord" in certain cases. Return only the exact name of the lessor in the answer_string field. Format any all-caps values in the text to title case. Write the response without using definitional terms. Refer to each entity or concept by its full name or a clear descriptor each time it is mentioned, rather than assigning it a shorthand label or nickname.

Context: {context}

Con esto, se obtuvo un accuracy del 0.64, superando la anterior variación.

Variación 8: Usar la variación 7 cambiando la temperatura a 0.1 del LLM

Hasta ahora, todos los experimentos realizados se hicieron fijando la temperatura de cero. En el contexto de los sistemas RAG, establecer la temperatura a cero tiene un propósito claro relacionado con el control de la aleatoriedad y la precisión en las respuestas generadas. Primero que todo, minimiza la aleatoriedad. La temperatura en los modelos de lenguaje controla el grado de aleatoriedad en la generación de texto. Con una temperatura de cero, el modelo siempre seleccionará el token (palabra o fragmento) con la probabilidad más alta en cada paso, asegurando que la respuesta sea lo más determinística posible. En aplicaciones donde la precisión es crucial (como consultas técnicas, legales o médicas), una temperatura baja elimina interpretaciones alternativas o creativas que podrían llevar a resultados incorrectos o incoherentes.

Sin embargo, se experimentó cambiando la temperatura hasta un valor máximo de 0.1 y usando la misma técnica utilizada en la variación 7.

Debido a que en esta aplicación la precisión es crucial y como no hubo mejoras al cambiar la temperatura, se descarta cambiar la temperatura de su valor de cero.

Variación 9: Utilizar LLM Listwise Rerank para la compresión contextual

LLM Chain Filters y LLM Listwise Rerank son dos enfoques utilizados para clasificar y refinar respuestas generadas o recuperadas por modelos de lenguaje (LLM). Ambos tienen como objetivo mejorar la relevancia y calidad de las respuestas, pero difieren significativamente en su enfoque y aplicación.

Los LLM Chain Filters se basan en la evaluación secuencial de una lista de posibles respuestas o documentos para filtrar aquellos que no cumplen con ciertos criterios predefinidos, como relevancia, precisión o claridad. En este proceso, cada elemento pasa a través de un modelo de lenguaje que decide si mantenerlo o descartarlo en función de una puntuación o criterio binario, como "apto" o "no apto". Este enfoque es particularmente útil para reducir rápidamente el número de elementos en una lista extensa, eliminando resultados irrelevantes y optimizando los recursos

computacionales en etapas posteriores. Sin embargo, una desventaja de este método es que podría descartar elementos valiosos si los filtros son demasiado estrictos, lo que limita la calidad de los resultados finales.

Por otro lado, los LLM Listwise Rerank reordenan una lista completa de respuestas o documentos asignando una puntuación relativa a cada elemento en comparación con los demás. Este enfoque considera la relación entre los elementos de la lista y genera un ranking optimizado en el que los elementos más relevantes o útiles se posicionan al principio. Este método se emplea cuando la lista inicial ya es de alta calidad, pero se requiere un orden más preciso o ajustado a las necesidades del usuario final. Aunque ofrece resultados más sofisticados al mantener todos los elementos en un orden optimizado, requiere mayor capacidad computacional y puede ser más lento, especialmente para listas largas (Ma et al., 2023).

Una diferencia clave entre ambos enfoques radica en cómo procesan la lista. Mientras que los LLM Chain Filters evalúan cada elemento de manera individual para decidir si eliminarlo, los LLM Listwise Rerank analizan la lista como un todo, teniendo en cuenta las relaciones contextuales entre los elementos. En términos de aplicación, los filtros suelen utilizarse en las primeras etapas del procesamiento, como una forma de preprocesamiento o eliminación rápida, mientras que el reranking es más adecuado para refinar y presentar los resultados finales (Ma et al., 2023).

Ambos métodos son complementarios y, a menudo, se combinan dentro de sistemas complejos para maximizar la relevancia y la eficiencia. Por ejemplo, se pueden aplicar filtros inicialmente para reducir una lista extensa y luego realizar un reranking sobre los elementos restantes para asegurar que los resultados finales sean de la mayor calidad posible. Sin embargo, una desventaja de LLM Listwise Rerank es que requiere mayor capacidad computacional, ya que evalúa la lista completa de forma conjunta por lo que puede ser más lento en comparación con los filtros, especialmente para listas largas (Ma et al., 2023).

En esta variación, sin embargo, se utilizó LLM Listwise Rerank de manera aislada para comprobar si el accuracy aumentaba.

Variación 10: Utilizar embedding filters

Realizar una llamada adicional al modelo de lenguaje para cada documento recuperado puede ser costoso y lento. Los embeddings filters ofrecen una alternativa más económica y eficiente, ya que utiliza embeddings para representar tanto los documentos como la consulta. Luego, solo devuelve aquellos documentos cuyas incrustaciones sean lo suficientemente similares a la consulta.

Para consultas que involucran múltiples criterios o requieren razonamiento lógico, los embeddings pueden no capturar adecuadamente todas las dimensiones necesarias, lo que puede reducir la precisión del filtro. Los embeddings generalmente se enfocan en el significado global de los textos y pueden no considerar aspectos estructurales o relacionales que podrían ser clave para determinar la relevancia de un documento en ciertos casos.

Los embeddings generalmente se enfocan en el significado global de los textos y pueden no considerar aspectos estructurales o relacionales que podrían ser clave para determinar la relevancia de un documento en ciertos casos.

Mejoras para la pregunta 2

Para la pregunta 2, se partió de la base que los LLM Chain Filters son la mejor opción para obtener el contexto relevante para la pregunta. Este es un resumen de los resultados:

Variación	Descripción	Validation accuracy	Test accuracy
1		0.3906	0.4667
2		0.3750	**0.5000**
3		0.3906	0.4333
4		0.3750	0.3333
5		0.4063	0.4333
6		0.3281	0.3333
7		0.3281	0.3000

Variación 1: Usar LLM Chain Filters con una pregunta de carácter general para obtener el contexto

En esta primera variación se utilizó la siguiente pregunta de contexto:

> What are all the dates in this document?

El prompt también se cambió de manera que sea específico para esta pregunta:

> On what date was the contract executed? This could the date the document was signed, or if that date is unavailable its the date on which the agreement was entered into or is dated. Provide the answer in the answer_date field in the format YYYY-MM-DD.
>
> Context: {context}

Variación 2: Usar la variación 1 cambiando la temperatura del LLM a 0.1

De igual forma que con la pregunta 1, se utilizó una temperatura de 0.1.

Variación 3: Utilizar un prompt diferente

Antes de experimentar con una pregunta de contexto, se experimentó utilizando un prompt diferente. Se utilizó el siguiente prompt:

> What is the effective date of the contract? In other words: on what date was the contract executed? This could be the date the document was signed, or if that date is unavailable it's the date on which the agreement was entered into or is dated. Provide the answer in the answer_date field in the format YYYY-MM-DD.
>
> Context: {context}

Variación 4: Utilizar CoT con una pregunta de contexto distinta

El prompt utilizado fue el siguiente:

> Please review the contract details provided and answer the following question:
>
> Question: What is the effective date of the contract?

Instructions:

1. Identify the signature date: Begin by checking the contract for a specific signature date. This is often listed near the end of the document, alongside the parties' signatures. If this date is present, it should generally be considered the effective date.

2. Check for an explicitly stated effective date: If the contract does not have a signature date or it is unclear, look for any explicitly stated "effective date" or "date of execution" within the contract text, which may sometimes be noted at the beginning or in an introductory section.

3. Determine fallback date: If neither a signature date nor an explicit effective date is provided, identify any other date that indicates when the agreement was entered into, such as a date mentioned in the title or heading.

4. Confirm and record the date: Once the effective date is identified, verify that it is in the correct format. Enter the answer in the `answer_date` field in the format YYYY-MM-DD.

Context: {context}

Y la pregunta de contexto fue la siguiente:

What is the effective date of the contract, or the date on which the contract was signed or entered into?

Variación 5: Utilizar CoT con la pregunta de contexto de la variación 1

Utilizando el mismo prompt de la variación 4 con la pregunta de contexto de la variación 1, se obtuvo un accuracy de 0.4043, por lo que se ratifica que lo que hace que esta métrica cambie es la calidad de la pregunta de contexto. En este caso, una pregunta más general (buscar todas las fechas en el documento), da resultados más relevantes que una pregunta específica, como la pregunta de contexto de la variación 4.

Variación 6: Utilizar el prompt de la variación 1 con la pregunta de contexto de la variación 4

Para ratificar la hipótesis de que la pregunta de contexto es el problema, se utilizó la pregunta de contexto específica de la variación 4 con el prompt de la variación 1.

Variación 7: Utilizar CoT en la pregunta de contexto

Debido a que se identificó que el problema es el contexto que se está obteniendo, se utilizó CoT en la pregunta de contexto. Se utilizó esta pregunta de contexto:

Question: What is the effective date of the contract?

Instructions:

1. Identify the signature date: Begin by checking the contract for a specific signature date. This is often listed near the end of the document, alongside the parties' signatures. If this date is present, it should generally be considered the effective date.

2. Check for an explicitly stated effective date: If the contract does not have a signature date or it is unclear, look for any explicitly stated "effective date" or "date of execution" within the contract text, which may sometimes be noted at the beginning or in an introductory section.

3. Determine fallback date: If neither a signature date nor an explicit effective date is provided, identify any other date that indicates when the agreement was entered into, such as a date mentioned in the title or heading.

Mejoras para la pregunta 3

En la pregunta 2, se evidenció que una pregunta de contexto de carácter más general dio mejores resultados. Por ende, se comenzó con este enfoque. Aquí un resumen de los resultados:

Variación	Descripción	Validation accuracy	Test accuracy
1	Utilizar LLM Chain Filters con una retriever question general	0.7541	0.7500
2	Utilizar la variación 1 con una temperatura de 0.1 en el LLM	0.7705	0.7500
3	Utilizar la variación 1 con LLM Listwise Rerank	0.7213	0.6786
4	Utilizar CoT en el prompt	0.7541	**0.7857**

Variación 1: Utilizar LLM Chain Filters con una pregunta de contexto de carácter general

Utilizando la siguiente pregunta de contexto:

What are all the dates in this document?

El prompt utilizado es el siguiente:

What is the lease commencement date? The commencement date is the date on which a lessor makes an underlying asset available for use by a lessee. Provide the answer in the answer_date field in the format YYYY-MM-DD.

Context: {context}

Variación 2: Utilizar la variación 1 con una temperatura de 0.1 en el LLM

Utilizando una temperatura de 0.1 se realizó el mismo experimento de la variación 1.

Variación 3: Utilizar LLM Listwise Rerank

Debido a que con la variación 1 ya se pudo obtener una mejora significativa, se trató de experimentar con LLM Listwise Rerank una vez más con los mismos parámetros de la variación 1. Sin embargo, se obtuvo un accuracy de 0.7191, empeorando el resultado.

Variación 4: Utilizar CoT en el prompt

El prompt utilizado fue el siguiente:

Please review the lease contract details provided and answer the following question:

Question: What is the lease commencement date?

Instructions:

1. Identify terms defining commencement: Begin by looking for a section in the contract that specifically defines the "commencement date." Often, this term will be mentioned early in the document or in clauses about lease duration or start of obligations.

2. Determine asset availability: Confirm if the date matches the time when the lessor makes the underlying asset available for use by the lessee. If there is any discrepancy or multiple dates are mentioned, prioritize the date that reflects actual access or readiness for use.

3. Check for any conditions or delays: If the commencement date is contingent on conditions or an event (e.g., completion of installation), ensure that the selected date is the one when these conditions are fulfilled.

4. Format the date: Once you identify the correct lease commencement date, enter it in the `answer_date` field in the format YYYY-MM-DD.

Context: {context}

Mejoras para la pregunta 4

A continuación se muestra el resumen de las variaciones y resultados para la pregunta 4:

Variación	Descripción	Validation accuracy	Test accuracy
1	Utilizar LLM Chain Filters con una pregunta de carácter general	0.5667	0.5714
2	Utilizar la variación 1 cambiando la temperatura del LLM a 0.1	0.5833	0.5714
3	Utilizar CoT en el prompt	0.5667	**0.6071**

Variación 1: Utilizar LLM Chain Filters con una pregunta de carácter general

Se utilizó la siguiente pregunta de contexto:

What are all the dates in this document?

Y el siguiente prompt:

According to the contract, when is the last day of the lease (i.e., the date on which the lessee will lose the right to use the assets)?. Provide the answer in the answer_date field in the format YYYY-MM-DD.

Context: {context}

Variación 2: Utilizar la variación 1 cambiando la temperatura del LLM a 0.1

Se intentó la misma variación 1 pero con una temperatura de 0.1.

Variación 3: Utilizar CoT en el prompt

El prompt utilizado fue el siguiente:

Please review the lease contract details provided and answer the following question:

Question: According to the contract, when is the last day of the lease (i.e., the date on which the lessee will lose the right to use the assets)?

Instructions:

1. Locate the lease term clause: Start by finding the section in the contract that specifies the lease duration. This often includes both the start and end dates or the total length of the lease term.

2. Identify any specific end date: Check if a specific end date is stated within the lease term clause. If so, confirm that this date aligns with the intended duration of the lease agreement.

3. Consider any conditions for early termination or extension: Review whether the lease includes any clauses about early termination or extension that could alter the original end date. Only provide the last day of the lease as specified in the contract, excluding any optional extensions unless stated as part of the term.

4. Format the date: Once you have identified the end date, enter it in the `answer_date` field in the format YYYY-MM-DD.

Context: {context}

Mejoras para la pregunta 5

A continuación se muestra el resumen de las variaciones y resultados para la pregunta 5:

Variación	Descripción	Validation accuracy	Test accuracy
1		0.9420	0.9667
2		0.9420	0.9667
3		0.9420	**0.9667**

Variación 1: Utilizar LLM Chain Filters con una pregunta de contexto de carácter general

Se utilizó la siguiente pregunta de contexto:

Find information about purchase

Y el siguiente prompt:

Does the lease contract grant the lessee the option to purchase any assets? Provide the answer in the answer_boolean field.

Context: {context}

Variación 2: Cambiar la temperatura del LLM a 0.1 con la variación 1

Nuevamente, se cambiaron los valores de temperatura a 0.1.

Variación 3: Utilizar CoT en el prompt

El prompt utilizado fue el siguiente:

Please review the lease contract details provided and answer the following question:

Question: According to the contract, when is the last day of the lease (i.e., the date on which the lessee will lose the right to use the assets)?

Instructions:

1. Locate the lease term clause: Start by finding the section in the contract that specifies the lease duration. This often includes both the start and end dates or the total length of the lease term.

2. Identify any specific end date: Check if a specific end date is stated within the lease term clause. If so, confirm that this date aligns with the intended duration of the lease agreement.

3. Consider any conditions for early termination or extension: Review whether the lease includes any clauses about early termination or extension that could alter the original end date. Only provide the last day of the lease as specified in the contract, excluding any optional extensions unless stated as part of the term.

4. Format the date: Once you have identified the end date, enter it in the `answer_date` field in the format YYYY-MM-DD.

Context: {context}

Mejoras para la pregunta 6

A continuación se muestra el resumen de las variaciones y resultados para la pregunta 6:

Variación	Descripción	Validation accuracy	Test accuracy
1		0.7794	0.7333
2		0.7794	**0.7767**
3		0.8971	0.7333

Variación 1: Utilizar LLM Chain Filters con una pregunta de contexto de carácter general

La pregunta de contexto para este caso es:

Find information about extending or renewing the lease

Y el prompt fue el siguiente:

Does the lease contract grant the lessee the option to extend or renew the lease beyond it stated term? Provide the answer in the answer_boolean field.

Context: {context}

Variación 2: Utilizar un valor de 0.1 para la temperatura del LLM

Se experimentó aumentando el valor de la temperatura del LLM a 0.1 con el mismo prompt y pregunta de contexto de la variación 1.

Variación 3: Utilizar CoT en el prompt

El prompt se modificó para utilizar la técnica de CoT y fue el siguiente

Please review the lease contract details provided and answer the following question:

Question: Does the lease contract grant the lessee the option to extend or renew the lease beyond its stated term?

Instructions:

1. Identify renewal or extension clauses: Begin by searching the contract for any section that mentions the terms "renewal," "extension," "option to extend," or "right to renew." This information is often located in clauses related to lease duration or terms for continuation.

2. Evaluate specific conditions: If an extension or renewal option is mentioned, examine any specific conditions that apply, such as required notice periods, rent adjustments, or other terms the lessee must meet to exercise this option.

3. Confirm that the option is granted to the lessee: Ensure the extension or renewal option specifically applies to the lessee (and not only to the lessor). Be aware of any restrictive clauses that might limit the lessee's ability to renew or extend.

4. Formulate answer: Based on your findings, enter `True` in the `answer_boolean` field if the lessee has the option to extend or renew the lease, and `False` if no such option is available.

Context: {context}

Mejoras para la pregunta 7

A continuación se muestra el resumen de las variaciones y resultados para la pregunta 6:

Variación 1: Usar LLM Chain Filters

La pregunta de contexto que se utilizó fue la siguiente:

What are the termination options for the lessee?

Y el prompt utilizado fue el siguiente:

Does the lease contract grant the lessee the option to terminate or cancel the lease before it stated term? Provide the answer in the answer_boolean field.

Context: {context}

Variación 2: Utilizar una pregunta de contexto más específica

Se utilizó una pregunta de contexto más específica para tratar de obtener contextos más relevantes. La pregunta de contexto fue la siguiente:

Find information about termination, expiration, cancellation, or early termination of the lease.

Variación 3: Utilizar una temperatura de 0.1 en el modelo LLM

Se utilizó una temperatura de 0.1 en el modelo LLM con los mismos parámetros de la varación 2.

Variación 4: Utilizar CoT en el prompt

Se utilizó la técnica de CoT en el prompt. El prompt fue el siguiente:

Please review the lease contract details provided and answer the following question:

Question: Does the lease contract grant the lessee the option to terminate or cancel the lease before the end of its stated term?

Instructions:

1. Identify termination clauses: Begin by reviewing the lease contract for any section that specifies termination or cancellation terms. Look for keywords like "early termination," "cancellation option," or "lessee right to terminate."

2. Evaluate conditions: If a termination clause is found, examine any specific conditions or requirements tied to this option. For example, check if it requires a notice period, payment of a penalty, or any other precondition.

3. Confirm lessee's rights: Determine if the contract clearly grants the lessee the right to terminate the lease before the end of its term. Distinguish between rights given exclusively to the lessee and those granted only to the lessor.

4. Formulate answer: Based on the findings, provide a boolean response in the `answer_boolean` field—`True` for 'Yes' if the lessee has the right to terminate early, and `False` for 'No' if they do not.

Context: {context}

Variación 5: Utilizar CoT con la pregunta de contexto de la variación 1

Utilizando la pregunta de contexto de la variación 1 y el prompt de la variación 4, se hizo un experimento obteniendo los siguientes resultados:

Mejoras para la pregunta 8

Variación	Descripción	Validation accuracy	Test accuracy
1	Utilizar LLM Chain Filters	0.9130	0.8965
2	Utilizar LLM Chain Filters con una pregunta de contexto más específica	0.9275	**0.9655**
3	Añadir más términos específicos a la pregunta de contexto	0.9710	0.8276
4	Cambiar la temperatura del LLM a 0.1	0.9710	0.8276

Variación 1: Utilizar LLM Chain Filters

En esta variación, se utilizó compresión contextual con LLM Chain Filters utilizando la siguiente pregunta de contexto:

What are the termination options for the lessor?

Y el siguiente prompt:

Does the lessor have the right to terminate or cancel the lease prior to the expiration date? Provide the answer in the answer_boolean field.

Context: {context}

Variación 2: Utilizar LLM Chain Filters con una pregunta de contexto más específica

Utilizando la siguiente pregunta de contexto:

What are the termination or expiration options?

Y usando el mismo prompt de la variación 1, se obtuvo los siguientes resultados:

Variación 3: Añadir más términos específicos a la pregunta de contexto

Se incluyeron más términos específicos a la pregunta de contexot, haciendo que sea la siguiente:

Find information about termination, expiration, cancellation, or early termination of the lease.

Variación 4: Cambiar la temperatura del LLM a 0.1

Utilizando los mismos parámetros de la variación 3, se cambió la temperatura del LLM a 0.1. Estos fueron los resultados:

Resumen de los mejores prompts

En la siguiente tabla, se resumen los mejores prompts por pregunta

Pregunta	Prompt
1	What is the name of the lessor in the agreement in the Context? The lessor is also referred to as the "landlord" in certain cases. Return only the exact name of the lessor in the answer_string field. Format any all-caps values in the text to title case. Write the response without using definitional terms. Refer to each entity or concept by its full name or a clear descriptor each time it is mentioned, rather than assigning it a shorthand label or nickname. Context: {context}
2	What is the effective date of the contract? In other words: on what date was the contract executed? This could be the date the document was signed, or if that date is unavailable it's the date on which the agreement was entered into or is dated. Provide the answer in the answer_date field in the format YYYY-MM-DD. Context: {context}
3	Please review the lease contract details provided and answer the following question: Question: What is the lease commencement date?

	Instructions: 1. Identify terms defining commencement: Begin by looking for a section in the contract that specifically defines the "commencement date." Often, this term will be mentioned early in the document or in clauses about lease duration or start of obligations. 2. Determine asset availability: Confirm if the date matches the time when the lessor makes the underlying asset available for use by the lessee. If there is any discrepancy or multiple dates are mentioned, prioritize the date that reflects actual access or readiness for use. 3. Check for any conditions or delays: If the commencement date is contingent on conditions or an event (e.g., completion of installation), ensure that the selected date is the one when these conditions are fulfilled. 4. Format the date: Once you identify the correct lease commencement date, enter it in the answer_date field in the format YYYY-MM-DD. Context: {context}
4	Please review the lease contract details provided and answer the following question: Question: According to the contract, when is the last day of the lease (i.e., the date on which the lessee will lose the right to use the assets)? Instructions: 1. Locate the lease term clause: Start by finding the section in the contract that specifies the lease duration. This often includes both the start and end dates or the total length of the lease term. 2. Identify any specific end date: Check if a specific end date is stated within the lease term clause. If so, confirm that this date aligns with the intended duration of the lease agreement. 3. Consider any conditions for early termination or extension: Review whether the lease includes any clauses about early

	termination or extension that could alter the original end date. Only provide the last day of the lease as specified in the contract, excluding any optional extensions unless stated as part of the term. 4. Format the date: Once you have identified the end date, enter it in the `answer_date` field in the format YYYY-MM-DD. Context: {context}
5	Please review the lease contract details provided and answer the following question: Question: According to the contract, when is the last day of the lease (i.e., the date on which the lessee will lose the right to use the assets)? Instructions: 1. Locate the lease term clause: Start by finding the section in the contract that specifies the lease duration. This often includes both the start and end dates or the total length of the lease term. 2. Identify any specific end date: Check if a specific end date is stated within the lease term clause. If so, confirm that this date aligns with the intended duration of the lease agreement. 3. Consider any conditions for early termination or extension: Review whether the lease includes any clauses about early termination or extension that could alter the original end date. Only provide the last day of the lease as specified in the contract, excluding any optional extensions unless stated as part of the term. 4. Format the date: Once you have identified the end date, enter it in the `answer_date` field in the format YYYY-MM-DD. Context: {context}
6	Does the lease contract grant the lessee the option to extend or renew the lease beyond it stated term? Provide the answer in the answer_boolean field. Context: {context}
7	Please review the lease contract details provided and answer the following question: Question: Does the lease contract grant the lessee the option to terminate or cancel the lease before the end of its stated term? Instructions:

	1. Identify termination clauses: Begin by reviewing the lease contract for any section that specifies termination or cancellation terms. Look for keywords like "early termination," "cancellation option," or "lessee right to terminate." 2. Evaluate conditions: If a termination clause is found, examine any specific conditions or requirements tied to this option. For example, check if it requires a notice period, payment of a penalty, or any other precondition. 3. Confirm lessee's rights: Determine if the contract clearly grants the lessee the right to terminate the lease before the end of its term. Distinguish between rights given exclusively to the lessee and those granted only to the lessor. 4. Formulate answer: Based on the findings, provide a boolean response in the `answer_boolean` field—`True` for 'Yes' if the lessee has the right to terminate early, and `False` for 'No' if they do not. Context: {context}
8	Does the lessor have the right to terminate or cancel the lease prior to the expiration date? Provide the answer in the answer_boolean field. Context: {context}

Despliegue

El despliegue del sistema es un paso crucial que involucra el diseño de la infraestructura adecuada, la gestión de certificados de seguridad y la automatización de procesos para garantizar la disponibilidad del sistema. En esta sección se detallan los tres enfoques explorados durante el proceso de despliegue del sistema AWS.

Enfoque con AWS Lambda y API Gateway

En un intento por hacer la solución más eficiente y escalable, se exploró la posibilidad de adoptar un enfoque serverless utilizando AWS Lambda y API Gateway. Esta arquitectura eliminaría la necesidad de administrar servidores, lo que reduciría los costos operativos y permitiría escalar automáticamente según la demanda. Sin embargo, surgió un obstáculo importante: muchas de las librerías de Python requeridas para la solución necesitaban la capacidad de multiprocesamiento (multiprocessing en inglés), la cual no es compatible con las funciones Lambda. Cualquier librería que requiera como subdependencia el módulo “multiprocessing” de Python tiene que tener un trato especial al ser usada en AWS Lambda. Un ejemplo es la librería “unstructured[pdf]”, que utiliza

multiprocesamiento para procesar PDFs. Esta limitación hizo que este enfoque fuera inviable para la solución propuesta.

Enfoque con Elastic Beanstalk y con una sola instancia

Se utilizó Elastic Beanstalk, pero sin un balanceador de carga (en AWS, este servicio se lo conoce como Elastic Load Balancer o ELB por sus siglas) para reducir costos. Para implementar la terminación SSL/TLS de manera gratuita, se utilizó el servicio de Let's Encrypt junto con el cliente "certbot" para obtener y renovar automáticamente los certificados SSL, dado que los certificados de AWS Certificate Manager no pueden descargarse utilizarlos en servidores EC2. Además, se implementó la automatización del proceso de Continuous Deployment utilizando GitHub Actions y el CLI de Elastic Beanstalk, lo que permitió desplegar cambios en la solución de manera rápida, automática y confiable. El diagrama de arquitectura propuesto para este enfoque es el siguiente:

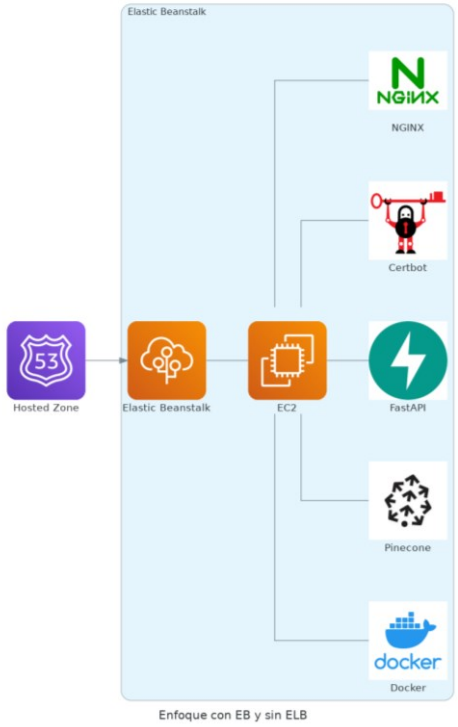

Figura 13 Infraestructura usando Elastic Beanstalk con solo una instancia EC2

Sin embargo, cuando se utilizó esta infraestructura, después de una semana y debido a la publicación de un artículo donde se mencionó a gaapRT en NYCA (https://fintechfundamentals.substack.com/p/the-future-of-finance-tech) como una empresa innovadora, el servidor empezó a recibir varios ataques DDoS, causando que el servidor se sobrecargara de peticiones que no podía responder y eventualmente causando interrupciones en el servicio.

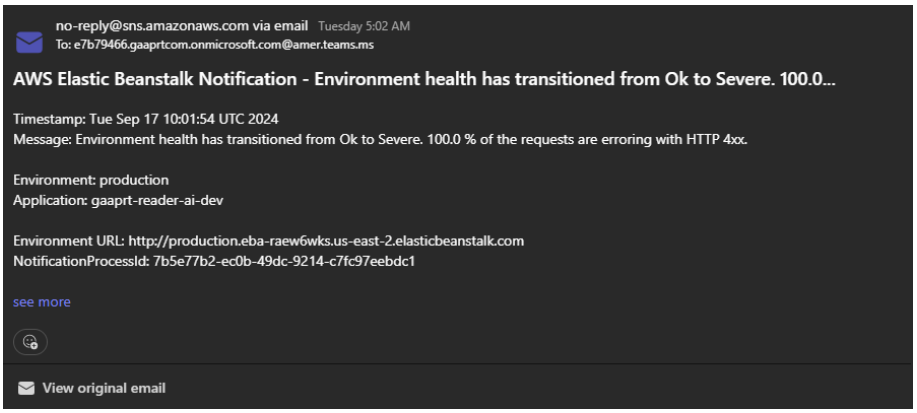

Figura 14 Notificación alertando que el servidor recibió muchas peticiones que devolvían un código 4xx debido a ataques DDoS

Se repensó la infraestructura para añadir una capa de seguridad extra por un costo adicional. La terminación SSL ahora se haría en un balanceador de carga y en frente del mismo se pondría un WAF (Web Application Firewall) para bloquear las peticiones maliciosas.

Enfoque con Elastic Beanstalk y un balanceador de carga

Se hicieron las modificaciones descritas anteriormente para terminar con la siguiente infraestructura:

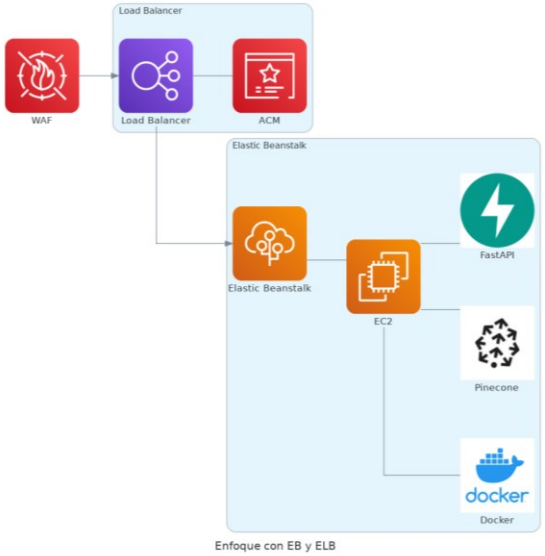

Figura 15 Infraestructura usando Elastic Beanstalk, un balanceador de carga y WAF para evitar peticiones maliciosas

Esta infraestructura detuvo los ataques DDoS y aparte dio visibilidad de qué peticiones maliciosas se estaban ejecutando y de dónde.

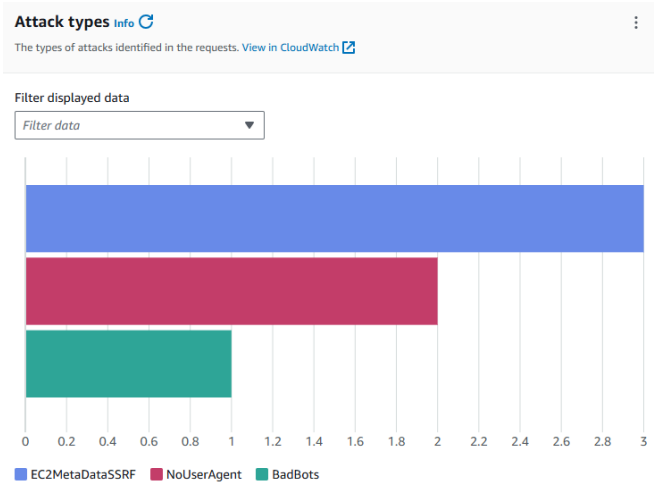

Figura 16 Ejemplo de tipo de ataques perpetrados

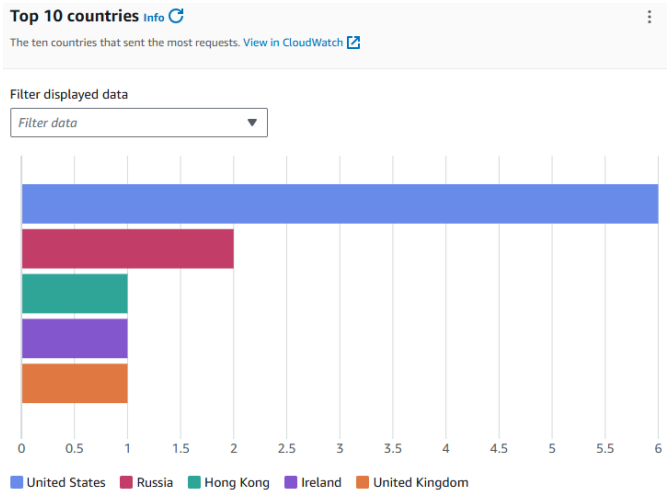

Figura 17 Ejemplo de países de donde vienen la mayoría de los ataques

Monitoreo y mejoras

A continuación se muestra un ejemplo de cómo la interfaz gráfica muestra la respuesta al usuario:

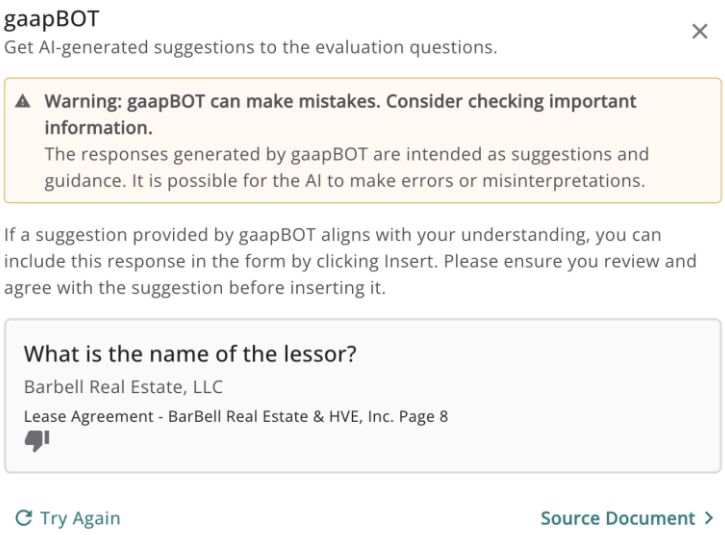

Figura 18 Ejemplo de respuesta para el usuario final

El monitoreo del sistema se llevó a cabo mediante un sistema de retroalimentación diseñado para evaluar la calidad de las respuestas generadas. Este sistema ofrece a los usuarios una forma para expresar su satisfacción o insatisfacción.

Para ello, se implementaron dos elementos principales:

1. **Opción de calificación directa:** Los usuarios podrán indicar que una respuesta no es buena utilizando un ícono de "manita hacia abajo".
2. **Registro de interacciones:** Se registrará si el usuario hace clic en las opciones "Try Again" o "Insert". Un clic en "Try Again" será interpretado como un indicador de que la respuesta inicial no cumplió con las expectativas, incluso si no se selecciona la "manita hacia abajo". Por otro lado, un clic en "Insert" reflejará que la respuesta fue considerada satisfactoria y aceptada.

Nota: El botón de “Try Again” regenerará la respuesta con una temperatura de 0.5.

Este enfoque permite identificar preguntas que requieren mejoras. Por ejemplo, las preguntas con un alto número de clics en "Try Again" o con una baja tasa de selección de "Insert" serán priorizadas para realizar ajustes en el proceso de generación de respuestas. Así, el sistema tendrá un ciclo de mejora continua basado en los datos obtenidos de los usuarios, asegurando la evolución constante de la calidad del sistema.

Trabajo futuro

Este trabajo de grado ha demostrado el potencial de los grandes modelos de lenguaje y técnicas como RAG para abordar desafíos específicos en el contexto de la normativa ASC 842. Sin embargo, aún quedan múltiples oportunidades de mejora que pueden explorarse en futuros desarrollos:

- **Exploración de otros grandes modelos de lenguaje**: Si bien este trabajo utilizó modelos de OpenAI como gpt-3.5-turbo, gpt-4o y gpt-4o-mini, en el futuro podrían evaluarse otros

modelos como los de Anthropic (Claude 3.5 Sonnet y Haiku), Google (Gemini 1.5 Flash y Pro), Meta (Llama 3.1 y 3.2) y las nuevas versiones de OpenAI (gpt-4o1 y gpt-4o1-mini). Estos modelos podrían ofrecer mejoras significativas en las métricas evaluadas.

- **Uso de modelos de código abierto**: Otra línea de mejora sería explorar modelos de código abierto para reducir costos asociados con licencias de modelos comerciales. Esta estrategia implicaría el despliegue en una infraestructura más compleja que requeriría GPUs.
- **Implementación de métricas adicionales**: Además de las métricas actuales como la exactitud y la distancia de Levenshtein, futuras métricas de evaluación podrían incluir:
    - **Relevancia**: Incorporar métricas como **Mean Reciprocal Rank (MRR)** y **Precision@K**, que evalúan la calidad de las respuestas y la relevancia de los documentos recuperados para el contexto.
    - **Latencia**: Medir el tiempo de respuesta del sistema sería crucial para ofrecer una mejor experiencia de usuario.
    - **Coverage**: Evaluar qué tan bien el sistema maneja una amplia variedad de consultas garantizaría que cubra todo el rango de preguntas posibles para la generación del memorando técnico contable en el contexto de la normativa ASC 842.
- **Ampliación de la muestra de datos**: Trabajar con una mayor cantidad y variedad de contratos de arrendamiento, incluyendo diferentes industrias, regiones y formatos, podría mejorar la capacidad del sistema para generalizar.
- **Automatización de la generación de memorandos**: El desarrollo de un sistema capaz de generar memorandos técnicos completos automáticamente, a partir de la información extraída, optimizaría aún más los procesos contables y reduciría la carga operativa de los profesionales contables.
- **Optimización sistemática del prompt**: Para mejorar la creación de prompts de manera sistemática, se podrían haber utilizado herramientas como AdalFlow o AutoPrompt. Estas herramientas permiten refinar los prompts de forma más estructurada y eficiente, lo que podría resultar en una mejora considerable en la calidad de las respuestas generadas por los modelos (Shin et al., 2020).

La implementación de estas mejoras no solo optimizaría el sistema actual, sino que también ampliaría su alcance, haciéndolo más robusto, adaptable y valioso en aplicaciones del sector contable.

Conclusiones

Este trabajo de grado ha permitido desarrollar un sistema basado en grandes modelos de lenguaje y técnicas de RAG para responder automáticamente preguntas en contratos de arrendamiento bajo la normativa ASC 842. El sistema diseñado ha sido evaluado de acuerdo con los objetivos propuestos y ha alcanzado resultados mixtos, algunos de los cuales superaron las expectativas establecidas (preguntas 5, 6, 7 y 8) y otros que evidencian áreas de mejora (preguntas 1, 2, 3 y 4).

Se logró desarrollar un sistema funcional para la extracción automática de información clave de los contratos de arrendamiento. Este sistema tiene como objetivo facilitar el trabajo de los profesionales contables al permitirles identificar entidades relevantes de manera precisa y eficiente. A lo largo del proceso, se alcanzó el objetivo general de lograr un sistema con un rendimiento adecuado, si bien se lograron variaciones en la precisión dependiendo de la pregunta. Con respecto a los objetivos específicos, se logró lo siguiente:

1. **Vectorización y almacenamiento de contratos**: El sistema fue capaz de cargar, procesar y almacenar los contratos en una base de datos vectorial, cumpliendo con los requisitos establecidos en el objetivo específico.

2. **Sistema de respuestas a preguntas cerradas**: Se diseñó un sistema que generó respuestas a preguntas basadas en la información extraída de los contratos. De las 8 preguntas planteadas, **5 superaron el umbral del 70%** de precisión, mientras que **3 no lo lograron**. Las preguntas que superaron el 70% fueron principalmente aquellas relacionadas con la existencia de opciones de compra, renovación y terminación anticipada del contrato. Sin embargo, las preguntas sobre fechas críticas como el inicio y la terminación del contrato mostraron un desempeño inferior al esperado, no alcanzando el 70% en precisión.

    - **Preguntas que superaron el 70%**:

        - Pregunta 5: ¿El arrendatario tiene la opción de comprar el objeto del arrendamiento?
        - Pregunta 6: ¿El arrendatario tiene la opción de renovar el contrato o extenderlo más allá de la fecha de terminación del contrato?
        - Pregunta 7: ¿El arrendatario tiene la opción de terminar el contrato de arrendamiento antes de la fecha de terminación del contrato?
        - Pregunta 8: ¿El arrendador tiene la opción de terminar el contrato de arrendamiento antes de la fecha de terminación del contrato?

    - **Preguntas que no superaron el 70%**:

        - Pregunta 1: ¿Cuál es el nombre del arrendador?
        - Pregunta 2: ¿En qué fecha se firmó el contrato?
        - Pregunta 3: ¿Cuál es la fecha de inicio del arrendamiento?
        - Pregunta 4: ¿Cuál es la fecha de terminación del contrato?

Estas variaciones en los resultados resaltan la necesidad de perfeccionar el modelo para preguntas que requieren extracción de datos más específicos o que involucran interpretaciones más complejas de los documentos.

3. **Estrategias de mejora**: En algunas de las preguntas que no superaron el 70%, la implementación de la estrategia de **Chain of Thought (CoT)**, que fomenta que el modelo realice un proceso de razonamiento más detallado antes de generar una respuesta, podría mejorar significativamente los resultados. El uso de CoT ayuda a los modelos de lenguaje a

pensar paso a paso, lo que podría ser particularmente útil para preguntas que requieren la interpretación de fechas o que involucren condiciones complejas en los contratos.

4. **Comparación de modelos de lenguaje**: La evaluación de modelos de diferentes proveedores, como OpenAI, demostró que algunos modelos más avanzados ofrecieron mejores resultados en términos de precisión. Sin embargo, los costos asociados con el uso de estos modelos fueron considerablemente más altos, lo que subraya la importancia de balancear el rendimiento con la eficiencia en aplicaciones comerciales.

5. **Despliegue y monitoreo del sistema**: El sistema se desplegó en una infraestructura basada en AWS, lo que permitió su integración con otros servicios y la implementación de medidas de seguridad como la encriptación de datos. Sin embargo, el monitoreo continuo y la retroalimentación del usuario serán esenciales para mejorar la calidad de las respuestas generadas y ajustar el modelo a las necesidades reales de los usuarios.

En conclusión, el sistema desarrollado logró superar los objetivos propuestos en gran medida, especialmente en cuanto a la precisión en la identificación de opciones de compra, renovación y terminación anticipada de contratos. Sin embargo, no todas las preguntas alcanzaron el umbral del 70%, lo que evidencia áreas de mejora. La implementación de estrategias como **Chain of Thought (CoT)** y el ajuste de modelos podrían ayudar a superar estos desafíos. Además, la experimentación con otros modelos de lenguaje y métricas adicionales como **Relevancia**, **latencia** y **cobertura** contribuirán a optimizar el sistema, mejorando su aplicabilidad y eficiencia en el sector contable.

Este trabajo fue exitoso en varias de las preguntas a resolver y ofrece una base sólida para futuras mejoras y adaptaciones, permitiendo que el sistema evolucione para satisfacer de manera más efectiva las necesidades del mercado y los profesionales contables.

Referencias

Accruent. (2023, May 8). *ASC 842 Lease Accounting*. Accruent. https://www.accruent.com/resources/blog-posts/asc-842-what-why-it-important

Amatriain, X. (2024). *Prompt Design and Engineering: Introduction and Advanced Methods*.

Amazon. (2024a). *¿Qué es una base de datos vectorial?* https://aws.amazon.com/es/what-is/vector-databases/

Amazon. (2024b). *What is Retrieval-Augmented Generation?* Amazon.

Crail, C., & Main, K. (2022, September 9). *Generally Accepted Accounting Principles (GAAP) Guide*. Forbes. https://www.forbes.com/advisor/business/generally-accepted-accounting-principles-gaap-guide/

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*.

Hugging Face. (2024). *Question Answering*. Hugging Face.

Jing, Z., Su, Y., & Han, Y. (2024). *When Large Language Models Meet Vector Databases: A Survey*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*.

Ma, X., Zhang, X., Pradeep, R., & Lin, J. (2023). *Zero-Shot Listwise Document Reranking with a Large Language Model*.

Merrit, R. (2023, November 15). *What Is Retrieval-Augmented Generation, aka RAG?* NVIDIA.

OpenAI. (2024a). *Models*. https://platform.openai.com/docs/models/gp

OpenAI. (2024b, January 25). *New embedding models and API updates*. https://openai.com/index/new-embedding-models-and-api-updates/

OpenAI. (2024c, January 25). *New embedding models and API updates*. https://openai.com/index/new-embedding-models-and-api-updates/

OpenAI. (2024d, August 6). *Introducing Structured Outputs in the API*. https://openai.com/index/introducing-structured-outputs-in-the-api/

OpenAI. (2024e, August 6). *Introducing Structured Outputs in the API*. https://openai.com/index/introducing-structured-outputs-in-the-api/

Prompt Engineering Guide. (2024a). *Chain-of-Thought Prompting*. https://www.promptingguide.ai/techniques/cot

Prompt Engineering Guide. (2024b). *Few-Shot Prompting*. https://www.promptingguide.ai/techniques/fewshot

Prompt Engineering Guide. (2024c). *Zero-Shot Prompting*. https://www.promptingguide.ai/techniques/zeroshot

PwC. (2024, January 31). *Leases*. https://viewpoint.pwc.com/dt/us/en/pwc/accounting_guides/leases/assets/pwcleasesguide0523.pdf

Saltz, J. (2024, April 28). *The GenAI Life Cycle*.

Shin, T., Razeghi, Y., Logan, R. L., Wallace, E., & Singh, S. (2020). *AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts*.

Sun, X., Ji, Y., Ma, B., & Li, X. (2023). *A Comparative Study between Full-Parameter and LoRA-based Fine-Tuning on Chinese Instruction Data for Instruction Following Large Language Model*.

Toloka. (2023, June 23). *The history, timeline, and future of LLMs*. Toloka. https://toloka.ai/blog/history-of-llms/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*.

Wang, Z. (2022). *Modern Question Answering Datasets and Benchmarks: A Survey*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*.

Worth, P. J. (2023). Word Embeddings and Semantic Spaces in Natural Language Processing. *International Journal of Intelligence Science*, *13*(01), 1–21. https://doi.org/10.4236/ijis.2023.131001

Anexos

Tabla 10 Respuestas a las preguntas de cada contrato

	Pregunta							
Contrato	1	2	3	4	5	6	7	8
1	BarBell Real Estate, LLC	2022-01-25	2022-01-25	2027-01-25	0	0	1	1
2	Sony Chocolate Industries Ltd	2024-03-01	2024-03-01	2029-03-31	0	1	1	1
3	The Wheelership LLC	2021-11-01			0	0	1	1
4	Venable Tenant, LLC	2015-06-20	2010-10-01	2021-07-31	0	1	1	1
5	Wyomissing Professional Center III Limited Partnership	1995-03-31	1995-04-01	2005-03-31	0	1	1	1
6	Hooten Non Exempt Family Trust B	2021-08-01	2021-08-01	2023-08-01	0	0	1	1
7	AR Industrial No. 1, Ltd	2009-02-10	2010-01-31	2020-01-31	0	1	1	1
8	Village Corner, LLC	2003-03-21			0	1	1	1
9	GH4 Partners LLC				0	0	0	1
10	YESCO Properties, LLC	2018-07-05	2018-02-01	2020-01-31	0	0	1	1
11	Holder Investments, Inc	2024-03-01	2024-03-01	2027-02-28	0	0	1	1
12	ORWIG PROPERTY MANAGEMENT CENTER SQUARE LLC	2015-09-01	2015-09-01	2025-09-01	1	1	1	1
13	Douglas R. Rippel	2017-01-28	2017-10-01	2022-09-30	0	1	1	1
14	Adairsville GA, LLC	2023-12-01	2023-12-01	2033-12-01	1	1	1	1
15	Narcoossee Acquisitions, LLC	2017-04-01	2017-04-01	2022-11-01	0	1	0	1
16	EXPONENT REALTY, LLC	2016-04-01	2016-04-01	2019-03-31	0	1	1	1
17	SRE TKC CHARLESTON IV, LLC	2021-01-26	2021-05-01	2026-08-01	0	1	0	1
18	BOYER RESEARCH PARK ASSOCIATES X, L.C.	2019-01-31			0	1	0	1
19	Chicago Executive Airport				0	0	1	1
20	UP 64 SIDNEY STREET, LLC	2020-10-01	2020-06-01	2032-02-29	0	1	1	1

21	INDUSTRIAL DEVELOPMENTS INTERNATIONAL, INC.	2007-04-27	2007-10-01	2017-10-31	0	0	1	1
22	ARE-MA REGION NO. 75, LLC	2021-12-20	2023-11-30	2035-02-01	0	1	1	1
23	MEADOWS OFFICE, L.L.C.	2007-10-09	2007-10-09	2017-10-09	0	1	1	1
24	APT Cowork, LLC				0	0	1	1
25	Wilks Ranch Texas, LTD	2022-11-01	2022-11-01	2033-11-01	0	1	1	1
26	ARE-708 QUINCE ORCHARD, LLC	2019-08-09	2019-09-01	2029-12-01	0		1	1
27	FARNAM STREET FINANCIAL, INC	2006-05-30			0	0	1	1
28	Industrias Asociadas Maquiladoras, S.A. de C.V.	2007-09-01	2007-09-01	2014-09-01	0	1	1	1
29	fischer group SE & Co. KG	2017-07-11	2021-09-01	2026-08-31	0	1	1	1
30	Eagle I Investments, L.L.C.	2016-06-20	2016-09-01	2019-08-31	0	1	1	1
31	CTROOK002 LLC	2022-09-01	2022-09-01	2037-09-01	1	1	0	1
32	ARE-SAN FRANCISCO NO. 63, LLC	2019-12-04	2020-11-01	2031-02-01	0	1	1	1
33	Landa Properties LLC	2022-01-31	2022-02-01	2023-01-31	0	1	0	1
34	Landa App LLC - 1703 Summerwoods Lane Griffin GA LLC	2022-02-01	2022-02-01	2023-01-31	0	1	0	1
35	Landa Properties LLC	2021-11-01	2021-11-01	2022-10-31	0	1	0	1
36	PINAL COUNTY	2022-08-10			1	0	1	1
37	Anthony Nigel Sampson	2021-07-15			0	0		
38	ARE-SD REGION NO. 71, LLC	2021-01-14	2021-12-01	2036-12-01	1	1	1	1
39	CARGILL, INCORPORATED	2021-09-24	2021-11-01	2026-10-31	1	0	1	1
40	BERNARDO WINDELL, LLC	2021-10-04	2021-12-01	2027-02-01	1	1	1	1
41	BERNARDO WINDELL, LLC	2021-10-04	2021-12-01	2027-02-01	0	1	1	1
42	100 CHESAPEAKE BLVD LLC	2021-08-10			1	1	0	1
43	Hudson Towers at Shore Center, LLC, a Delaware limited liability company	2021-04-16	2021-06-01	2023-06-30	0	0	1	1
44	Taylyn Holdings, LLC	2023-07-01	2023-06-01	2033-06-01	1	1	1	1

45	MINOAN CAPITAL PTY LTD	2020-01-29	2020-02-01	2023-01-31	0	0	0	1
46	UPPER CHEASPEAKE FLEX ONE, LLC	2021-08-10	2021-12-01	2028-12-01	0	1	1	1
47	Apple Moving, Inc.	2019-11-01	2019-11-01	2025-01-01	1	1	0	1
48	Sunrise Nominee Trust	2022-10-01	2022-11-01	2027-10-31	0	0	1	1
49	Cruiser Lane, LLC	2012-12-13	2021-02-01	2019-01-31	1	1	0	1
50	NWP BUILDING 20 LLC	2021-09-01			0	0	0	1
51	Phil Bosua	2022-09-22	2021-09-01	2023-09-30	0	0	1	1
52	VISIONS FEDERAL CREDIT UNION	2022-12-07	2023-01-01	2024-12-31	0	1	1	1
53	Hooten Non Exempt Family Trust B	2021-08-01	2021-08-01	2023-08-01	0	0	1	1
54	Trustmark National Bank	2021-03-17	2021-06-01	2028-08-01	0	1	0	1
55	Price-Poore House, LLC	2022-09-28	2022-10-01	2023-12-31	0	0	1	1
56	State of Wyoming	2017-02-02	2019-08-02	2029-08-01	0	0	0	1
57	Queen Mary Bioenterprises Limited	2022-10-20	2022-10-20	2027-10-19	0	0	1	1
58	Sichuan Anyi Hengke Technology Co., Ltd.	2021-09-01	2021-09-01	2026-02-28	0	1	1	1
59	GIFFORD INVESTMENTS, INC.	2006-05-19	2006-03-01	2016-02-29	1	0	1	1
60	Kathryn Joy Atkinson	2022-04-01	2022-04-01	2027-04-01	0	1	1	1
61	Overlook At Rob Roy Owner, LLC	2020-09-18	2020-10-01	2020-12-31	0	0	1	1
62	Phil Bosua	2022-09-22	2021-09-01	2023-09-30	0	0	1	1
63	Hangzhou Zhexin Information Technology Co., LTD	2020-08-26	2020-09-11	2022-10-05	0	1	1	1
64	Beijing Guochuan Borui Technology Co., Ltd.	2021-01-18	2021-01-18	2024-01-17	0	1	1	1
65	Beijing Guochuan Borui Technology Co., Ltd.	2021-01-18	2021-01-18	2022-01-17	0	1	1	1
66	STATE OF WYOMING	2017-02-02	2019-08-02	2029-08-01	0	0	0	1
67	McCLELLAN FARM	2013-08-07	2013-08-07	2023-08-07	0	1	1	1

68	STATE OF WYOMING	2017-02-02	2019-11-02	2029-11-01	0	0	0	1
69	105 W. First Street Owner, L.L.C.	2020-07-24	2022-03-24					
70	Mundo Talio SL	2021-02-23	2020-09-01	2021-07-01	0	0	1	1
71	Steven and Janet Atkinson	2024-04-01	2024-04-01	2029-04-01	0	1	1	1
72	Kathryn Joy Atkinson	2024-04-01	2024-04-01	2029-04-01	0	1	1	1
73	ALC Aircraft Limited	2023-02-14	2023-03-01	2028-03-01	0	0	0	1
74	5550 Nicollet, LLC	2019-12-16	2020-01-01	2020-12-31	0	1	1	1
75	HP LUMINA, LLC	2015-04-26	2015-04-26	2022-04-30	1	0	1	1
76	TARGET GROUP INC.	2019-12-20	2018-06-28	2019-06-28	0	0	0	0
77	MALONE US ROUTE 2 WATERBURY PROPERTIES, LLC	2015-10-19	2016-05-01	2026-05-01	0	1	1	1
78	Rehco Holdings, LLC		2021-12-01	2026-12-01	0	0	1	1
79	Ever Winland Limited	2021-02-26	2020-12-15	2022-10-02	1	0	0	0
80	Ailanthus L.L.C.	2019-10-31	2019-11-01	2021-10-31	0	1	1	1
81	VT Aviation Leasing LLC	2021-12-21	2021-12-21	2022-12-21	0	0	0	1
82	VT Equipment Leasing LLC	2021-12-21	2021-12-21	2022-12-21	0	0	0	1
83	Rehco Holdings, LLC		2021-12-01	2026-12-01	0	0	1	1
84	MALONE US ROUTE 2 WATERBURY PROPERTIES, LLC	2015-10-19	2016-05-01	2026-05-01	0	1	1	1
85	HP LUMINA, LLC	2015-04-26	2015-04-26	2022-04-30	1	0	1	1
86	ALC Aircraft Limited	2023-02-14	2023-03-01	2028-03-01	0	0	0	1
87	ALC Aircraft Limited	2023-02-14	2023-03-01	2028-03-01	0	0	0	1
88	Steven and Janet Atkinso	2022-04-01	2022-04-01	2027-04-01	0	1	1	1
89	Queen Mary Bioenterprises Limited		2022-10-20	2027-10-19	0	0	1	1
90	Ailanthus L.L.C.	2019-10-31	2019-11-01	2021-10-31	0	1	1	1

91	WESTWIND ACQUISITION COMPANY, L.L.C.	2020-12-22	2020-12-22	2021-12-22	0	0	1	1
92	Ever Winland Limited	2021-02-26	2020-12-15	2022-10-02	1	0	0	1
93	Penta Partners, LLC	2018-01-29	2018-01-01	2023-06-30	0	1	1	1
94	THE COUNCIL OF THE CITY OF MANCHESTER	1988-02-08	1988-12-01	2138-12-01	0	0	0	0
95	Mo Industripark AS	2021-07-19	2021-08-12	2031-08-11	0	1	1	1
96	UCB, INC.	2020-01-01	2020-01-01	2022-12-31	0	0	1	1
97	CARGILL, INCORPORATED	2021-09-24	2021-11-01	2026-10-31	1	0	1	1
98	Beijing Hontao Management Consulting Co., Ltd.	2020-03-30	2020-03-20	2022-03-29	0	1	1	1
99	Wickfield Phoenix LLC	2020-01-09	2020-02-01	2025-01-31	0	0	1	1
100	McCLELLAN FARM	2013-08-07	2013-08-07	2023-08-07	0	1	1	1