

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO INGENIERIA COMERCIAL
SANTIAGO - CHILE



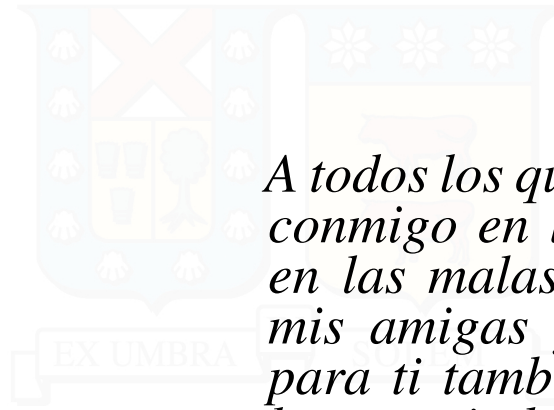
**RIESGOS ASOCIADOS A LA CREACIÓN Y USO DE APLICACIONES UTILIZANDO
MODELOS GRANDES DE LENGUAJE (LLM)**

SANTIAGO JESÚS VASCONCELLO ACUÑA

TESIS PARA OPTAR AL TÍTULO DE
INGENIERO COMERCIAL

PROFESOR GUÍA : SR. PABLO ISLA
PROFESOR CORREFERENTE : SR. THIERRY DE SAINT PIERRE.

Diciembre 2023



*A todos los que estuvieron
conmigo en las buenas y
en las malas, mi familia,
mis amigas y amigos. Y
para ti también, para mí
hoy me titulo por los dos
...*

AGRADECIMIENTOS

Aunque resulta casi imposible el agradecer a todos los que han estado conmigo en esta epopeya que ha sido salir de la universidad, es necesario hacer el intento, porque no ser agradecido tanto del apoyo que he recibido, además de ser inconsciente, sería extremadamente presumido. Todos somos el promedio de las personas que frecuentamos y yo definitivamente he tenido mucha suerte.

En ámbito laboral, muchísimas gracias Varder Johnson, en Chaski e Ignacio Quintana, en Moneda Asset Managment, quienes confiaron en mí y me permitieron ser sus colegas. También a Esteban Róman, Claudio Orrego y Lucas Vallejo, que fueron mis compañeros de trabajo, de los cuales aprendí mucho más de lo que alguna vez imaginé.

Y aunque pueda resultar atípico, también agradezco a las comunidades online que formo parte, especialmente a la comunidad de Platzi y de en_coders, que además de ayudar cuando se necesita, también brindan compañía en el solitario camino del estudio y aprendizaje autodidacta.

En mi vida universitaria, le agradezco a los profesores: Fernando Díaz, María Pía Santibáñez, Francisca González, Rodrigo Ortega, Joaquín Dagnino y Norman Dabner por permitirme ser su ayudante y aprender mucho de ellos. Y a Thierry de Saint-Pierre, por confiar de manera tan ciega en dos estudiantes y permitirnos trabajar en proyectos que - hasta ese momento - no sabía que podían realizarse.

Por otro lado, no puedo dejar de mencionar a quienes alimentaron mi curiosidad y me alentaron a estudiar. Agradezco a Jorge Bayer, por enseñarme a programar y cambiarme la vida desde el primer script. A Mario Vega, el mejor profesor de matemáticas que existe y que además de ayudarme a confiar en mí, me enseñó lo hermosas que pueden llegar a ser las matemáticas.

Gracias a mis amigos del CMAD: Juan, Benjamín, Constanza, María Olga, Bastián, Anaís y Jorge. Gracias por compartir conmigo lo nerd que puedo llegar a ser. A mis amigas: Paulina, Gabriela, Isidora, Ximena, Valeria y Yanira, gracias por estar ahí para mí, por ser mis compañeras de trabajo y mi sostén emocional en la universidad.

Un especial agradecimiento, a Ignacio Cortes, Marco Figueroa y Marcelo Escudero, gracias por ser mis mejores amigos de toda la vida, los quiero mucho y no los cambiaría por nada.

A mis abuelos, siempre atentos de cómo iba y en especial a ti mamita, vuelvo a agradecer y pido perdón por no alcanzar a titularme antes de que partieras, espero que si existe otro plano puedas ver que lo logre. También a Rayén y mi sobrina Tamara, atentas y presentes. Gracias también a mi tía Luchita, tío Marco, Eli, Benjamín y José Pablo, porque a veces la familia no solo es de sangre.

Para terminar, siempre estará mi familia. Mi madre que ha estado en las buenas y en las malas, siempre en eterna vigilia para apoyarme cuando lo necesitaba, no sería nada si ti mamá. A mi padre, siempre presente, siempre agradeceré la paciencia y la oportunidad que me diste de equivocarme para ser lo que soy ahora.

Mi hermana, mi apoyo en todo momento, confidente y la base emocional en mis peores momentos, sin ti no hubiese ni siquiera podido entrar en la universidad. Son lo más importante en mi vida, siempre faltarán palabras para agradecerles. Y finalmente, para terminar un ciclo más, quiero decir que esto también va por ti Cristóbal, porque, aunque el destino nos separó, siempre vivirás en mí. Te dedico este trabajo, porque para mí hoy nos titulamos los dos.

Muchísimas gracias a todos los que son parte de mi vida, son mi improbabilidad matemática favorita.

RESUMEN EJECUTIVO

Este trabajo investiga los riesgos inherentes a la creación y uso de aplicaciones basadas en Modelos Grandes de Lenguaje (LLM) en la industria. Su foco es en Procesamiento del Lenguaje Natural (NLP) específicamente la generación de texto, excluyendo otras formas de inteligencia artificial generativa. El proyecto se basa en la experiencia de desarrollar una aplicación utilizando LLM y analiza los riesgos que pueden afectar tanto al equipo de desarrollo como a los resultados obtenidos. Se determinan los factores de riesgo en la creación y uso de estas aplicaciones, utilizando como caso de estudio un proyecto de búsqueda de jurisprudencia en tribunales ambientales. La metodología empleada incluye la creación del proyecto, un ejemplo práctico de uso y la evaluación de riesgos en cada etapa del proceso. El objetivo es proporcionar una estructura para aplicaciones que usen LLM, analizar los problemas y riesgos asociados con el uso de información para alimentar estos modelos, incluyendo un proceso completo de ETL (Extract, Transform, Load).

Palabras Clave. Modelos Grandes de Lenguaje (LLM), Generación de Texto, Riesgo, Inteligencia Artificial, Proceso de ETL, Jurisprudencia Ambiental

ABSTRACT

This work investigates the inherent risks in the creation and use of applications based on Large Language Models (LLM) in the industry. Its focus is on Natural Language Processing (NLP), specifically text generation, excluding other forms of generative artificial intelligence. The project is based on the experience of developing an application using LLM and analyzes the risks that can affect both the development team and the obtained results. The risk factors in the creation and use of these applications are determined, using as a case study a jurisprudence search project in environmental courts. The methodology employed includes the creation of the project, a practical example of use, and the evaluation of risks at each stage of the process. The goal is to provide a structure for applications that use LLM, analyze the problems and risks associated with the use of information to feed these models, including a complete ETL (Extract, Transform, Load) process.

Keywords. Large Language Models (LLM), Text Generation, Risk, Artificial Intelligence, ETL Process, Environmental Jurisprudence

Índice de Contenidos

1. Introducción	1
1.1. Problemática	2
1.2. Objetivos	2
1.2.1. Contexto	2
1.2.2. Objetivo General	2
1.2.3. Objetivos Específicos	2
1.3. Metodología	3
1.3.1. Estado del arte	3
1.3.2. Creación del Proyecto	3
1.3.3. Resultados	4
1.3.4. Evaluación de Riesgos	4
2. Estado del Arte	5
3. Creación del Proyecto	9
3.1. ETL	10
3.1.1. Extract	11
3.1.2. Transform	13
3.1.3. Load	16
3.2. Chatbot	18
3.3. Funcionamiento básico	20
4. Resultados	23
4.1. Caso 1: Consultar por resumen	23
4.2. Caso 2: Consulta por listado	24
4.3. Caso 3: Comparación entre reclamaciones	25
4.4. Caso 4: Consultar por información falsa	26
4.5. Reflexiones generales de los resultados	27
5. Evaluación de Riesgos	28
5.1. Creación del Proyecto	28
5.1.1. No tener un análisis previo de que se busca lograr	28
5.1.2. Calidad de los datos	29
5.1.3. Sesgos	29
5.1.4. Elección correcta del modelo	29
5.1.5. Costos Monetarios	30
5.1.6. Funciones de Embedding	30
5.1.7. Conocimiento de Framework	30
5.1.8. Volatilidad del Mercado	31
5.2. Uso de la Aplicación	32
5.2.1. Alucinaciones	32
5.2.2. Entrega de contexto adecuado	32

5.2.3. Limitaciones de la similitud de cosenos	32
5.2.4. Uso de información privada	33
5.2.5. Aprendizaje por Refuerzo con Retroalimentación Humana (RLHF)	33
6. Conclusiones	34
7. Recomendaciones	35
A. Contexto enviado por el Chatbot	40



Índice de Figuras

2.1. Estructura del perceptrón	5
2.2. Screenshot de la pagina de ChatGPT	6
2.3. Línea de tiempo del Transformer	7
2.4. Diagrama de la arquitectura un Transformer	7
2.5. Diagrama del funcionamiento de un Embedding	8
3.1. Estructura basica de la aplicación	9
3.2. Estructura del proceso de ETL para el Buscador Ambiental	10
3.3. Screenshot del Buscador de la pagina del Primer Tribunal Ambiental	11
3.4. Screenshot de una sola reclamación en la pagina web del buscador ambiental	11
3.5. Diagrama de secuencia para el proceso de extracción (Extract) de datos del proyecto	12
3.6. Estructura de carpetas y datos del proyecto	13
3.7. Diagrama de funcionamiento de la función mapreduce en un documento	14
3.8. Diagrama de secuencia para el proceso de transformación (Trasform) de datos	15
3.9. Diagrama de funcionamiento de ChormaDB	16
3.10. Diagrama de secuencia para el proceso de carga (Load) de datos	17
3.11. Diagrama de funcionamiento del Chatbot	18
3.12. Diagrama en consulta simple	19
3.13. Diagrama en consulta mediante Retriver augmented generation	19
3.14. Screenshot del Chatbot desarrollado	20
3.15. Representación vectorial de un prompt luego de pasar por la función de embedding	21
3.16. Screenshot de un jupyter notebook representando el proceso interno de selección	21
3.17. Screenshot del funcionamiento del Chatbot	22

1 | Introducción

La inteligencia artificial, también conocida como IA, ha experimentado un notable auge en la industria en los últimos tiempos de la mano de la llamada Industria 4.0 [11] y dentro del imaginario colectivo gracias a sus aplicaciones y fácil acceso, especialmente en ámbitos que durante mucho tiempo fueron algo reacios al cambio, como la administración y las finanzas [10]. Este incremento no se debe necesariamente a un aumento en la capacidad de cómputo a la que tenemos acceso, ya que esta capacidad ha ido creciendo gradualmente a lo largo del tiempo. Dicho lo anterior, aunque muy importante e investigado, la inteligencia artificial no generaba tanto interés como en la actualidad, porque esta estaba reservada para investigadores e implementaciones dentro de diversas industrias. No fue sino hasta que la empresa OpenAI lanzó la que es hasta hoy su producto estrella. El 30 de noviembre de 2022 fue el día que el público general pudo experimentar, probar y comprender de manera más completa la gran revolución llamada inteligencia artificial generativa con la entrada a las masas de ChatGPT [35], implantando en la población el concepto de inteligencia artificial generativa.

De acuerdo con Google, “La inteligencia artificial generativa se refiere al uso de la IA para crear contenido, como texto, imágenes, música, audio y video” [1]. Ahora gracias a una interfaz amigable para todo público, resultó sencillo para personas de diversas industrias descubrir que existían diversas herramientas capaces de generar texto y responder preguntas de manera comprensible, incluso para aquellos que no eran expertos en el uso de este tipo de tecnologías y con ello también dar paso a la popularización de otros tipos de aplicaciones en donde la inteligencia artificial generativa también era prometedora, como podría ser la generación de imágenes.

La génesis de esta tesis se basa en la experiencia de llevar a cabo un proyecto utilizando las tecnologías descritas previamente, para que a partir de ella se pueda discutir sobre los riesgos asociados tanto de la creación como del uso de estas tecnologías. Bajo este contexto, entendemos el riesgo como la probabilidad de eventos adversos y su impacto debido al uso y desarrollo de sistemas de IA, abarcando desde fallas técnicas, ataques maliciosos hasta desafíos éticos y efectos socioeconómicos imprevistos [11]. Este riesgo se considera desde el desarrollo, hasta el uso de la aplicación creada.

El proyecto se centró en el uso de uno de los grandes modelos de lenguaje, conocido como LLM por sus siglas en inglés, limitándose a la su capacidad de generación de texto. Por lo tanto, no profundizaremos en otros tipos de inteligencia artificial generativa, como la generación de imágenes o audio. El enfoque principal de esta tesis se concentra solo en el área del procesamiento del lenguaje natural aplicados a LLM.

1.1. Problemática

En Chile, se desarrollan numerosos proyectos en muchos sectores como pueden ser: la minería, inmobiliario, agrícola, salmonero, etc. Sin embargo, se ha observado un incremento en las reclamaciones sobre temas ambientales asociadas a estos proyectos, convirtiendo la “Permisología ambiental” en una problemática significativa para la generación de proyectos de toda índole dentro del país.

Muchos proyectos de inversión están judicializados, con sus causas llevadas ante el tribunal ambiental, debido a ello es crucial para las empresas poder prever estas reclamaciones y demandas potenciales con la mayor precisión posible [54][6][19]. Cada proyecto que se retrasa o cancela incrementa los costos y el riesgo, afectando no solo al proyecto en sí, sino también la percepción del país como destino amistoso para inversiones extranjeras.

Además es necesario brindar a los directivos o personal pertinente un acceso inmediato y comprensible a información relevante de estos tribunales ambientales, superando las barreras del lenguaje legal que puede resultar complejo para quienes no son expertos en derecho. Esto les permitirá tener una visión general sobre posibles acciones, consecuencias y problemas que podrían surgir al tomar cualquier tipo de decisiones.

1.2. Objetivos

1.2.1. Contexto

Esta tesis se desarrolló en conjunto con la empresa NAVIGO (www.navigo-ai.com) quien planteó como desafío el desarrollo de un proyecto haciendo uso de inteligencia artificial. Se utilizaría Inteligencia Artificial Generativa, términos a aclarar en el capítulo 2, específicamente modelos grandes del lenguaje (LLM) para abordar aplicaciones empresariales que resuelvan problemas difíciles de abordar sin el uso de este tipo de tecnologías. En particular esto se da cuando la información disponible no es estructurada, como es el caso de las sentencias jurídicas. La propuesta fue generar una herramienta que solucionara este problema, para que sin necesidad de un conocimiento muy robusto en leyes, fuera posible interpretar y tomar decisiones con mayor velocidad a las sentencias jurídicas de los Tribunales Ambientales de Chile.

1.2.2. Objetivo General

Reconocer los factores de riesgo asociados; tanto con la creación, como con el uso de aplicaciones que implementan Modelos Grandes de Lenguaje (LLM) en la industria, usando como caso de estudio el proyecto de búsqueda de jurisprudencia en tribunales ambientales.

1.2.3. Objetivos Específicos

1. Estructurar u organizar (un o el) contexto sobre inteligencia artificial, en específico la inteligencia artificial generativa y el procesamiento de lenguaje natural, que aborda esta Tesis.
2. Analizar el estado del arte en lo que refiere a Modelos Grandes del Lenguaje y sus conceptos asociados.

3. Desarrollar un proceso de ETL pertinente a la creación de la aplicación.
4. Explicar el desarrollo de la aplicación asociada al chatbot, analizando los resultados obtenidos de este mismo.
5. Evaluar sobre los resultados obtenidos por el chatbot.
6. Analizar y explicar los riesgos que conlleva la creación y uso de aplicaciones que utilizan Modelos Grandes de Lenguaje.
7. Analizar y concluir sobre cómo mitigar el riesgo que tienen este tipo de tecnologías.

1.3. Metodología

La metodología empleada en esta tesis se estructura en torno a cuatro componentes esenciales: el desarrollo del estado del arte, la creación del proyecto, los resultados asociados al proyecto y la evaluación de los riesgos asociados a cada etapa del proceso, tanto en la fase de desarrollo del proyecto como en la aplicación práctica del mismo.

1.3.1. Estado del arte

El Estado del Arte dará contexto sobre la inteligencia artificial, con un enfoque en los Modelos de Lenguaje de Gran Escala (LLM) y aplicaciones como ChatGPT. Se discutirá la evolución y el impacto de la inteligencia artificial, especialmente en el procesamiento del lenguaje natural (NLP) y la generación de contenido. También se explorará la arquitectura Transformer, clave para entender la eficacia de ChatGPT y otros modelos similares, y se abordará el proceso de “embedding” en el contexto de los modelos de lenguaje, entre otros.

1.3.2. Creación del Proyecto

Este capítulo aborda el desarrollo de un chatbot para consultar jurisprudencia en los Tribunales Ambientales de Chile, utilizando datos del “Buscador Ambiental” y la tecnología GPT-4 de OpenAI. Los pasos en su desarrollo son los siguientes:

- **ETL - Extracción, Transformación y Carga:** Proceso clave para la preparación de datos.
 - **Extracción:** Uso de Selenium y API para obtener datos del 'Buscador Ambiental'.
 - **Transformación:** Conversión de PDF a TXT, y aplicación de técnicas de map-reduce.
 - **Carga:** Almacenamiento de datos procesados en ChromaDB.
- **Desarrollo del Chatbot:** Creación del chatbot utilizando Python, Flask y FastAPI.
 - **Frontend:** Interfaz de usuario para realizar consultas.
 - **Backend:** Integración con LangChain, uso de estructura RAG ,OpenAI GPT-4 para procesamiento y respuesta.

1.3.3. Resultados

En esta sección de la tesis, se evaluará el desempeño del chatbot creado previamente, se estudiará su funcionamiento y los resultados observando si estos coinciden con las indicaciones que se les entregó de manera previa. Se presentarán cuatro casos de estudio para analizar la calidad de respuestas según los diferentes prompts que se entregarán, abordando desde resúmenes de casos judiciales hasta solicitudes de información inexistente, lo cual revelará tanto las fortalezas como las limitaciones del sistema, especialmente en su capacidad para manejar la precisión y veracidad de la información.

1.3.4. Evaluación de Riesgos

Para terminar, previamente, se trabajará en la identificación y análisis de los potenciales riesgos en cada una de las etapas del proceso, así como los riesgos derivados de los resultados obtenidos en cada uno de los tipos de consulta. Esta evaluación de riesgo se estructura de la siguiente manera:

- **Riesgos en la Creación del Proyecto:** Identificando los posibles obstáculos y contratiempos que pueden aparecer durante todas las etapas del desarrollo de una aplicación o herramienta que trabaje bajo los mismos tópicos que el chatbot realizado en esta tesis.
- **Riesgos en el Uso de la Aplicación:** Considerando de los riesgos asociados a la implementación práctica de proyectos con tecnologías similares a las desarrolladas en esta tesis.

Esta metodología proporciona un enfoque integral para la creación y aplicación de un proyecto utilizando grandes modelos de lenguaje (LLM), permitiendo una evaluación de los riesgos en cada etapa del proceso y en el caso específico de uso del chatbot creado para esta tesis. Esto facilita la toma de decisiones informadas y la formulación de estrategias para mitigar posibles contratiempos y dificultades previamente no estudiadas.

2 | Estado del Arte

Actualmente, el desarrollo y uso de la inteligencia artificial está en boca de todos, pero ¿qué es esta famosa inteligencia artificial y por qué ha adquirido tanta relevancia recientemente? A pesar de que ha sido desafiante definir este concepto durante mucho tiempo, podemos decir que la inteligencia artificial, también conocida como inteligencia de máquina (Machine Learning en inglés), es el uso de la inteligencia demostrada por la tecnología y máquinas [25]. En general, la inteligencia artificial, abreviada como AI (del inglés Artificial Intelligence) o IA (de la palabra en español), engloba técnicas como el aprendizaje automático, el aprendizaje profundo y otros aspectos de la inteligencia artificial [25]. Estos temas no son nuevos y han sido objeto de estudio durante muchos años. Por ejemplo, en el caso del aprendizaje profundo (Deep Learning en inglés), este se basa en el perceptrón, que fue descubierto en 1958 [49], por lo que no es un descubrimiento reciente. No fue hasta nuestros tiempos, cuando el poder de cómputo y las interfaces han sido democratizadas para el común de los usuarios, que hemos podido experimentar y entender realmente lo que la inteligencia artificial puede realizar por nosotros.

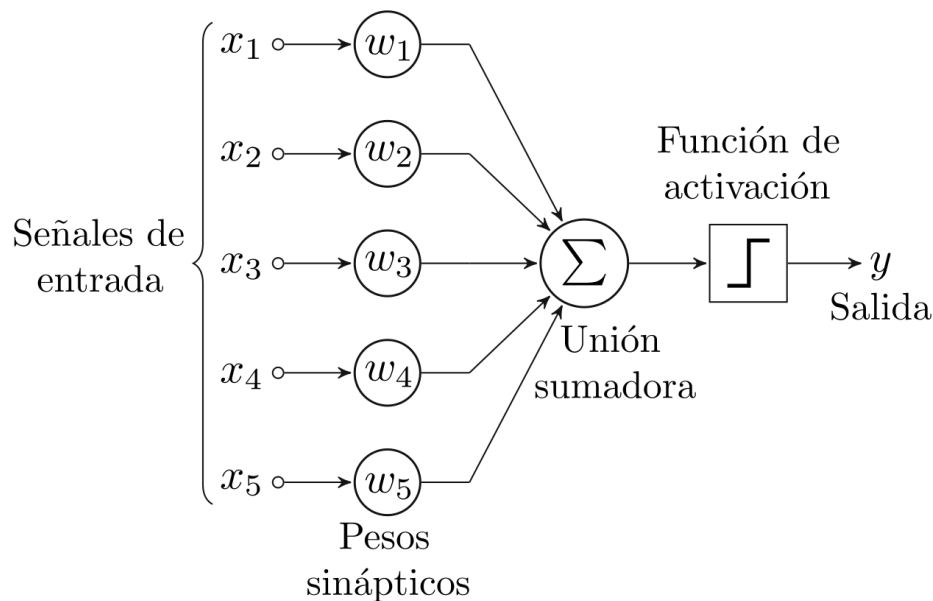


Figura 2.1: Diagrama de un perceptrón,
(Fuente: Wikipedia)

Luego de entender lo que es la inteligencia artificial queda la pregunta ¿Qué es lo diferente actualmente que no era capaz de ofrecer en el pasado? El 30 de noviembre de 2022 fue abierto al público la aplicación ChatGPT por la empresa especialista en inteligencia artificial OpenAI [42], deslumbrando a todos con la capacidad de responder las preguntas que los usuarios les entregaban, elevando aún más el interés por esta empresa y por este tipo de tecnología. La base de esta herramienta nace de una rama específica de la inteligencia artificial llamada procesamiento del lenguaje natural, abreviado NLP, en si siglas en inglés Natural Language Processing, siendo este un subcampo de la Inteligencia Artificial y lingüístico, dedicado a hacer que las computadoras comprendan declaraciones o palabras escritas en lenguajes humanos [13], llevando así el estudio de no solo texto sino de la lingüística y la sintaxis que los textos contienen.

El verdadero impacto que provocó ChatGPT en el mundo, fue el conocimiento popular de lo que hoy llamamos inteligencia artificial generativa, que puede ser definida como una técnica de inteligencia artificial que genera artefactos sintéticos analizando ejemplos de entrenamiento; aprendiendo sus patrones y distribución; y luego creando facsímiles realistas. La inteligencia artificial generativa (GAI) utiliza la modelización generativa y los avances en el aprendizaje profundo (DL) para producir contenido diverso a gran escala utilizando medios existentes como texto, gráficos, audio y video [27]. Por lo que, el público en general pudo entender que existían herramientas que podían crear diversos contenidos y con ello el boom de estas tecnologías entre la población fue cada vez más grande.



Figura 2.2: Screenshot de la pagina de ChatGPT

(Fuente: ChatGPT)

Cuando hablamos de inteligencia artificial y sus aplicaciones, usualmente nos referimos a la generación de diferentes modelos. Sin embargo, en el contexto de la inteligencia artificial generativa, como las herramientas que producen texto o asisten en problemas de Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés), también estamos hablando de un modelo de inteligencia artificial. La diferencia radica en que estos son considerablemente más grandes en términos del volumen de datos que manejan. Dado que están enfocados en temas de lenguaje, comúnmente los denominamos modelos grandes de lenguaje o LLM, por sus siglas en inglés de “Large Language Models”, siendo estos formalmente definidos como herramientas de inteligencia artificial (AI) basadas en redes neuronales recurrentes multicapa que son entrenadas con vastas cantidades de datos para generar texto similar al humano [3].

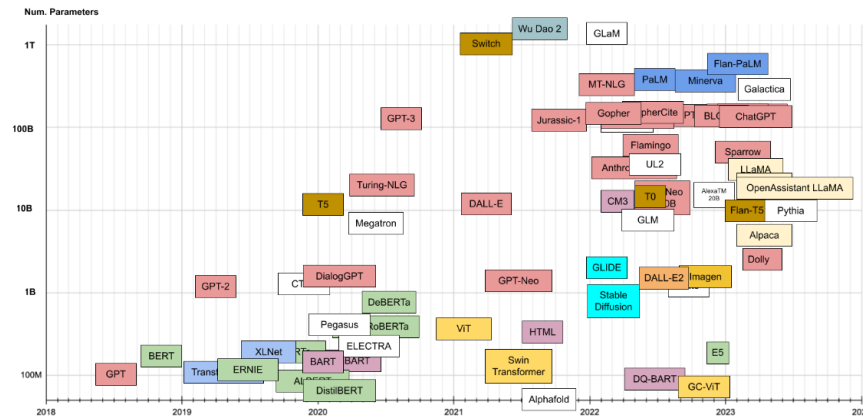


Figura 2.3: Línea de tiempo del Transformer. En el eje vertical, número de parámetros. Los colores describen la familia Transformer, (Fuente: Transformer models: an introduction and catalog [5])

ChatGPT funciona mediante una arquitectura base llamada Transformer [56], arquitectura creada por Google, que ha generado la gran revolución en la inteligencia artificial como la conocemos hasta la fecha, debido a que no necesariamente se centra en NLP, sino en inteligencia artificial generativa en general. Si somos aún más específicos, GPT viene de Transformer generativo pre entrenado, “Generative Pre-trained Transformer” en inglés, siendo esta una arquitectura con habilidad de comprender el lenguaje de mejor manera usando la arquitectura de los Transformers [48]. Aunque no fue hasta que este modelo creció hasta una cierta cantidad de parámetros que pudo demostrar sus capacidades en la gran amalgama de tareas que existen en el procesamiento de lenguaje natural, incluyendo la generación de texto como tal [8].

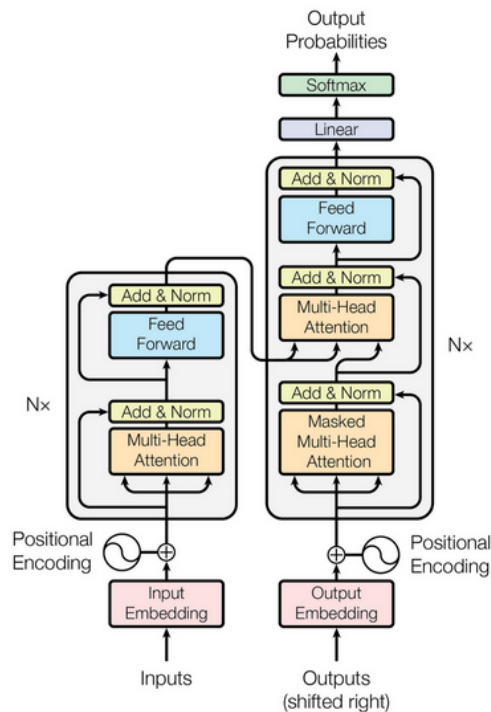


Figura 2.4: Diagrama de la arquitectura un Transformer, (Fuente: Attention Is All You Need [56])

Complementado lo anterior, es importante entender el funcionamiento de estos modelos ¿Cómo es posible que logren entender lo que escribo? Los modelos GTP funcionan en base a redes neuronales las cuales no entienden ni de letra y palabras, por lo que este texto tiene que pasar por una función de Embedding. Embedding es el proceso en el que representamos un texto, párrafo o documento de manera numérica, siendo esta representación en vector de múltiples dimensiones [40], estos vectores se pueden “graficar” en un espacio multi dimensional y con ello es posible ver la cercanía de cada uno de estos vectores entre ellos, por lo que este vector sirve como punto de entrada para el funcionamiento de los modelos GTP [39]. Además, se suelen usar para tareas como búsqueda, agrupación, recomendaciones, búsqueda de anomalías, etc [47].

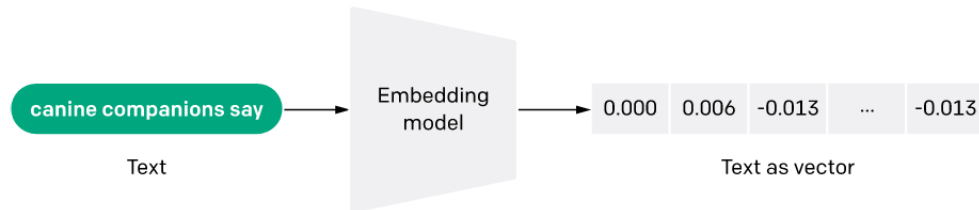


Figura 2.5: Diagrama del funcionamiento de un Embedding,
(Fuente: OpenAI[41])

Finalmente, todos estos conceptos nos ayudarán a entender el trasfondo de todas estas tecnologías y como estas interactúan con un Chatbot ¿Pero qué es un Chatbot? Para ello, primero será necesario entender que denominamos prompt, a pesar de que no existe una definición como tal, podríamos decir que son instrucciones textuales y ejemplo de interacciones deseadas que se antepone a las entradas de los modelos grandes de lenguaje, estos pueden sesgar los modelos y generar salidas deseadas, mejorando así la calidad de las salidas que generan estos modelos [62], esto es importante porque dependiendo que y como es la pregunta, o input, que se da a un modelo grande de lenguaje este puede tener una mejor o peor respuesta. Estos prompt son importantes porque son lo que recibe el chatbot, que esta potenciado por los LLM, por lo que podemos definir un chatbot como un programa informático diseñado para simular conversaciones con usuarios humanos [2], especialmente a través de Internet, que será el proyecto que acompaña a esta tesis.

3 | Creación del Proyecto

En Chile, los Tribunales Ambientales son órganos jurisdiccionales especiales, sujetos a la superintendencia directiva, correccional y económica de la Corte Suprema, cuya función es resolver las controversias medioambientales de su competencia y ocuparse de los demás asuntos que la ley somete a su conocimiento [33]. Estos tribunales generan una gran cantidad de jurisprudencia debido a los procesos como demandas, reclamaciones, etc. Que ellos atienden continuamente y que a su vez pueden ser encontrada en su portal de consulta llamado buscador ambiental [9].

Esta tesis contempla un proyecto consistente en la generación de un chatbot, donde se pueda preguntar sobre la jurisprudencia de los tribunales anteriormente descritos, por parte de cualquier persona sin necesidad de contar con una mucha expertiz en el ámbito legal. Sin embargo, por razones de capacidad de cómputo y limitaciones de la tecnología el chatbot se verá acotado en sus capacidades a solamente procesar las reclamaciones recibidas por el tribunal y no otro tipo de litigios que pueda atender dicho tribunal. Para este chatbot, se consideró una estructura como la representada en la Figura 3.1.

El proyecto esta contruido por diferentes etapas que consisten en: un proceso de extracción de datos desde el buscador ambiental, otro de transformación de estos datos para su utilización, la generación de vectores de los datos transformados para que puedan interactuar con la aplicación, una carga de estos vectores en una base de datos, para que después el chatbot que se creará pueda interactuar con ellos y mandara esa información al gran modelo de lenguaje (LLM), siendo en este caso el modelo “gpt-4” perteneciente a OpenAI, siendo la estrutura basica del proyecto la de Figura 3.1.

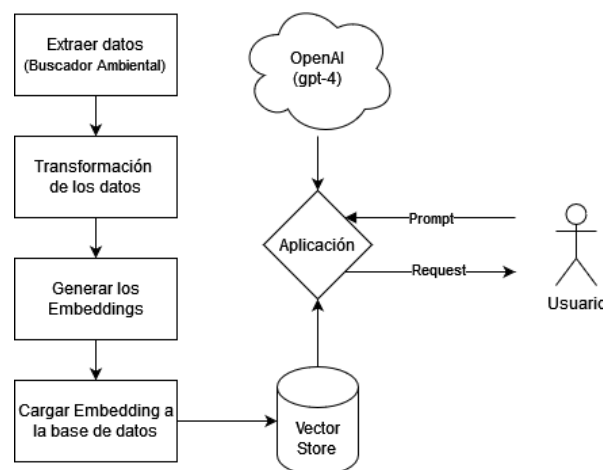


Figura 3.1: Estructura basica de la aplicación
(Fuente: Elaboración propia)

A partir de esta estructura básica anteriormente mencionada en la Figura 3.1, se explicará parte por parte el proceso con mucha más profundidad para entender tanto el funcionamiento como el proceso de creación, por lo que para empezar se iniciará con el proceso de ETL.

3.1. ETL

Para comenzar con la realización del proyecto, fue necesario realizar un proceso de ETL. El término ETL se refiere a las técnicas de "Extracción, Transformación y Carga" (Extract, Transform, Load), que constituyen un proceso clave para los datos necesarios para el proyecto. Este proceso implica la extracción de datos de fuentes heterogéneas, su transformación para ajustarse a las necesidades del negocio y su posterior carga en un destino que, por lo general, es un almacén de datos diseñado para el análisis y la generación de informes o aplicaciones [4]. Siendo en este proyecto el proceso como el que se muestra en la Figura 3.2.

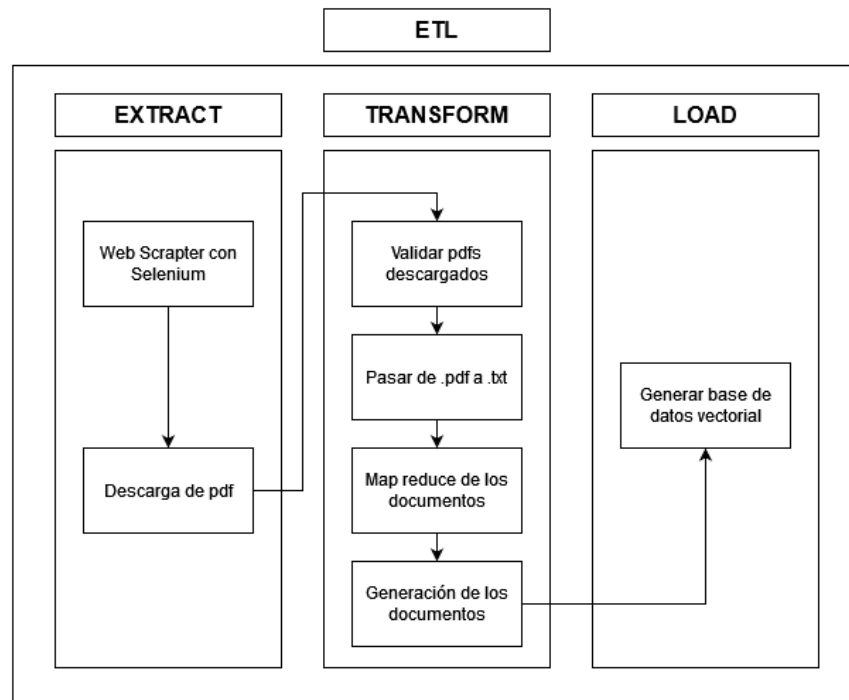


Figura 3.2: Estructura del proceso de ETL para el Buscador Ambiental

(Fuente: Elaboración propia)

Para la fase de extracción de los datos, esta implica la recolección de datos de múltiples fuentes, que pueden variar desde bases de datos estructuradas hasta información no estructurada en la web u otra fuente. La transformación se refiere al proceso de limpieza, conversión, y consolidación de estos datos en un formato adecuado para el análisis o entendimiento de la aplicación que se usará. Finalmente, la carga es el proceso de transferir los datos transformados al sistema de destino, donde se pueden almacenar y utilizar para la toma de decisiones estratégicas [4].

3.1.1. Extract

Iniciando la extracción de los datos, la información requerida para el desarrollo del Chatbot se obtuvo del “Buscador ambiental” del Tribunal de Protección Ambiental de Chile a través de su sitio web [9]. Este portal aloja todos los documentos públicos disponibles para su consulta en cualquiera de los tres tribunales ambientales. Para acceder a la base de datos necesaria y extraer el contenido necesario, se llevó a cabo la creación de un bot, definiendo un bot como un programa de software automático, repetitivo y con tareas predefinidas [28], capaz de recopilar de manera automática cada una de las entradas de este buscador, de manera análoga a cualquier usuario convencional que consultaría una jurisprudencia en el buscador.

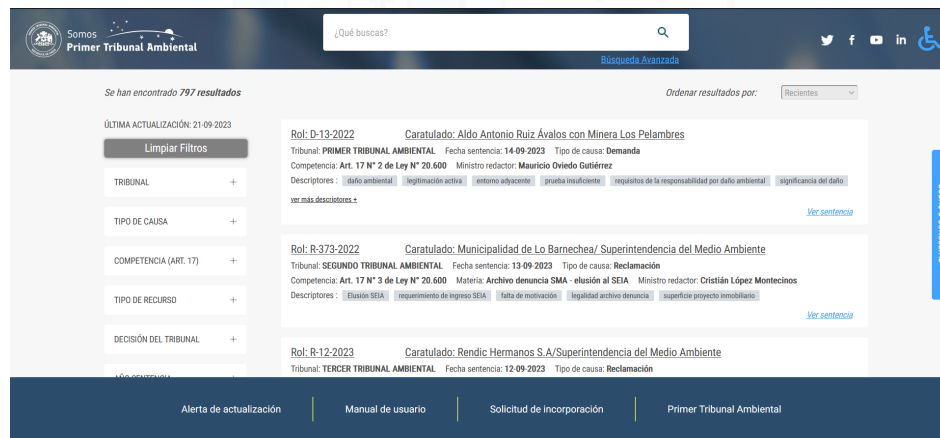


Figura 3.3: Buscador de la pagina del Primer Tribunal Ambiental
(Fuente: Pagina del Primer Tribunal Ambiental)

Para esta tarea, se empleó Selenium [50], una herramienta originalmente diseñada para generar pruebas unitarias en software, pero que, debido a la naturaleza reactiva y dinámica de los sitios web, así como la activa detección de bots por parte de algunas páginas, resulta ser la elección más apropiada para la extracción de datos del buscador. Este bot, después de explorar cada una las páginas del buscador ambiental, como se ilustra en la Figura 3.3, logró recolectar cada uno de los enlaces individuales que conducen a las páginas específicas de cada caso, tal como se muestra en la Figura 3.4, para de ellas luego extraer más información.



Figura 3.4: Screenshot de una sola reclamación en la pagina web del buscador ambiental
(Fuente: Pagina del Primer Tribunal Ambiental)

Posteriormente, se contempló la posibilidad de obtener tanto los enlaces a cada documento en formato PDF como la información detallada de cada uno de estos documentos, pudiendo llamar a esta información metadata, mediante la creación de un nuevo bot. Sin embargo, durante el proceso de desarrollo de este bot, se logró utilizar el mismo bot anteriormente usado, pero modificandolo, por lo que al hacerlo permitía obtener todos los datos mencionados anteriormente. Esto suprimió la necesidad de crear otro tipo de bot utilizando Selenium, ya que esta modificación logró obtener dichos resultados.

Para completar la fase de extracción de datos (Extract), una vez que se había obtenido toda la información mediante este bot con Selenium, el último paso consistió en generar un script, definiendo script como una serie de instrucciones que son interpretadas por un programa [58], que generará nuevas solicitudes con el objetivo de descargar todos los archivos PDF de cada una de las entradas del buscador. Como una limitación a la extracción de datos se decidió solo descargar las reclamaciones e ignorar el resto de los documentos que recibe el tribunal ambiental, cosa de acotar el proyecto en si y a limitaciones de capacidad de computo que se disponía. Estos archivos luego de ser descargados, estan listos para la próxima etapa del proyecto, que implica la transformación de los datos con el fin de procesar esta información, que luego será necesaria para construir la base de datos a partir de los documentos descargados. Dejando todo el proceso de “Extract” como se muestra en la Figura 3.5.

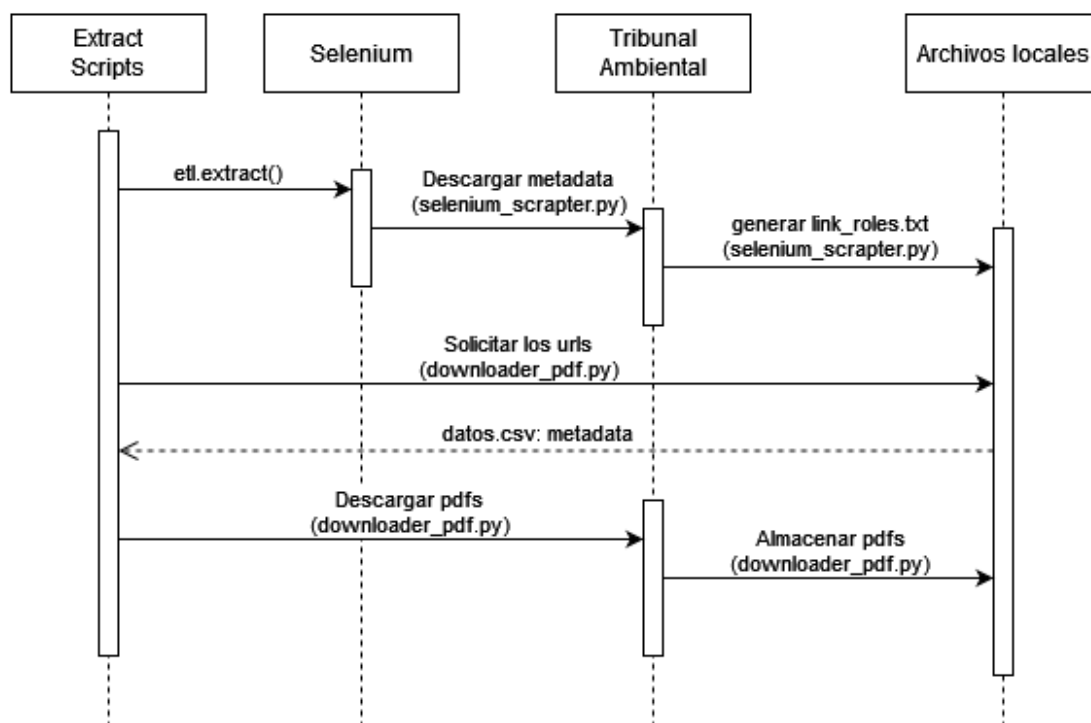


Figura 3.5: Diagrama de secuencia para el proceso de extracción (Extract) de datos del proyecto

(Fuente: Elaboración propia)

3.1.2. Transform

Continuando con el proceso de ETL, los PDFs que previamente habían sido descargados en el proceso de Extract, ahora requieren ser sometidos a modificaciones en la estructura y formato de sus datos, con el objetivo de convertir la información que inicialmente se presenta en un estado “sucio” a uno “limpio” cosa que puedan ser usados adecuadamente dentro de las posteriores tareas del proyecto. A esta serie de actividades de modificaciones a los datos lo denominamos “Transformación”, o “Transform,” en inglés.

Entre los datos descargados, nos encontramos con un extenso número de PDFs que presentan una serie de dificultades significativas para su manipulación dentro del programa. Esto se debe a que el Tribunal Ambiental no sigue un formato estándar en la estructura de las reclamaciones presentadas. En consecuencia, cada uno de los textos posee un formato propio, lo que complica en gran medida la extracción eficiente de las diversas secciones contenidas en dichos textos. Sin embargo, gracias al funcionamiento del proceso de semejanza semántica que se empleara por el chatbot, esta diversidad de formatos no representa un problema insuperable para el proyecto.

No obstante, surgen dificultades adicionales cuando se trata de las reclamaciones que son presentadas a los tribunales ambientales en formato digital o, en su defecto, en forma de fotocopias. Esto implica que no todos los documentos están habilitados para su procesamiento y con ello la extracción de su información. En consecuencia, el primer paso en el proceso de transformación involucra la discriminación de qué PDFs son susceptibles de ser procesados y cuáles no. Para llevar a cabo esta tarea, se ha desarrollado un script capaz de detectar texto dentro de un archivo PDF. Si el texto es legible, se almacena; de lo contrario, se elimina de los archivos descargados.

Una vez separados los PDFs legibles y adecuados para el trabajo posterior, se procedió con la transformación de estos documentos al formato TXT (texto plano). Esta etapa se realizó, considerando la conveniencia de trabajar con archivos en formato de texto en comparación con los archivos en formato PDF puro, dado que el próximo método de transformación, que implica el uso de map-reduce utilizados por Langchain, requiere que los datos estén en formato que sea interpretable por el framework.

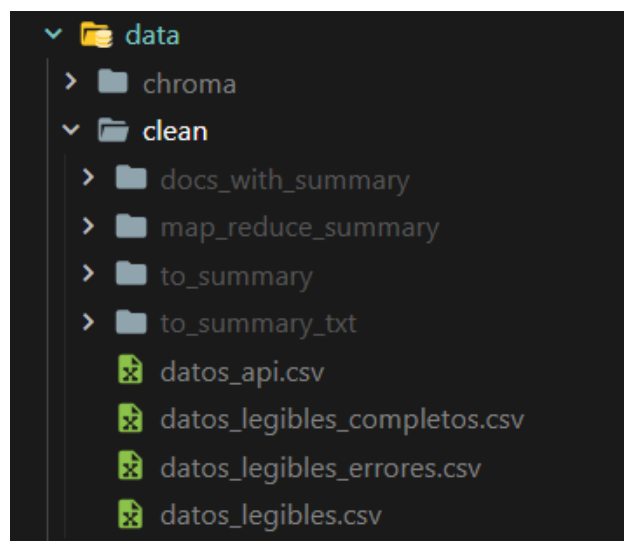


Figura 3.6: Estructura de carpetas y datos del proyecto]

(Fuente: Elaboración propia)

El funcionamiento interno de tanto el proceso de ETL como el futuro Chatbot, se basa en el framework Langchain para la interacción con el Modelo de Lenguaje Grande (LLM). LangChain es un framework poderoso que simplifica el desarrollo de aplicaciones utilizando modelos de lenguaje grandes (LLM). Proporciona una interfaz única y personalizable capaz de gestionar diferentes LLM, incluida la gestión rápida, el procesamiento, el aumento de datos, la orquestación de agentes, el almacenamiento y la evaluación. Este marco versátil permite a los desarrolladores integrar perfectamente los LLM con sus flujos de trabajo del mundo real y datos con el mínimo esfuerzo [52].

El proceso de Map-Reduce llevado a cabo por Langchain, es un modelo de programación diseñado para procesar grandes cantidades de datos de manera eficiente, escalable y distribuida a través de clústeres de servidores. En el contexto de un archivo PDF muy extenso, por ejemplo, si se requiere resumir el contenido o analizar la frecuencia de ciertas palabras, Map-Reduce podría ser utilizado para dividir la tarea en partes más pequeñas y manejables. Siguiendo con el ejemplo del PDF extenso en primera instancia, la función de map tomaría el texto del PDF y lo dividiría en elementos más pequeños, como párrafos o líneas, asignando a cada uno un resumen intermedio [38]. Luego, la función reduce recogería todos los resúmenes intermedios asociados con el documento extenso y los combinaría para producir un resultado agregado, con un resumen de todo el documento.

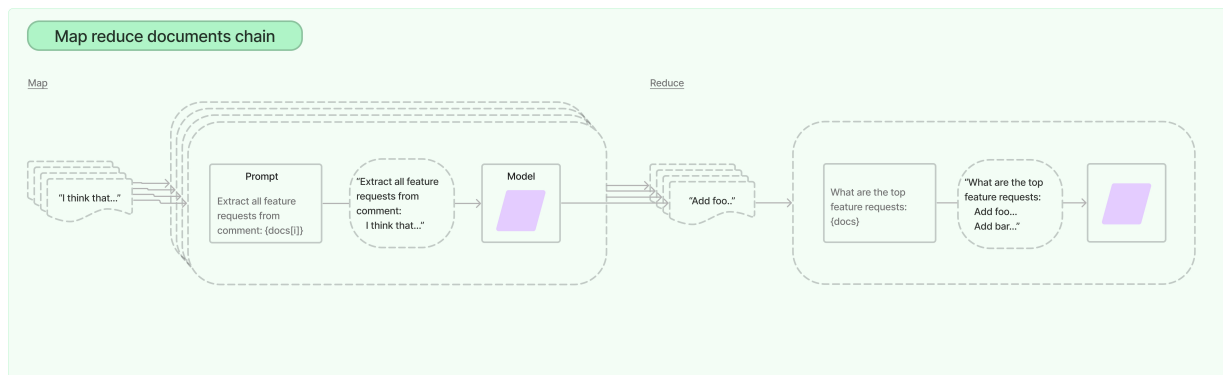


Figura 3.7: Diagrama de funcionamiento de la función mapreduce en un documento

(Fuente: Langchain [12])

Continuando con el proceso de “transform” y entrando a un ‘sub-proceso de map-reduce, es importante destacar que un archivo .txt obtenido del proceso anterior, puede contener una extensión demasiado grande, lo que conlleva a que tenga número de tokens muy elevado como para ser reducido de manera inmediata. En situaciones de este tipo, es necesario recurrir a un proceso de subdivisión que fragmenten los textos en segmentos con un número de tokens inferior al límite impuesto por la API OpenAI a los modelos de LLM de la empresa, siendo en este caso el modelo “gpt-4” que tiene un límite de 8,192 tokens por llamada a la API [45]. Cada archivo .txt puede ser dividido, resumido y exportado a un nuevo archivo .txt una vez que ha sido fragmentado previamente en segmentos.

Los documentos procesados son combinados utilizando otro proceso por parte de Langchain, con el fin de obtener un resultado final con todos los datos de los documentos consolidados. Para concluir el proceso de transformación, los resúmenes generados después de haber pasado por el procedimiento de map-reduce se someten a una última tarea, antes de ser incorporados en la base de datos mediante el proceso de “load”. Este paso implica la fusión de los resúmenes con la información obtenida a través de la información extraída por Selenium previamente y que es específica de cada texto llamada metadata, la cual es añadida a cada documento final en forma de texto. Este proceso resulta en la creación de un único documento para cada reclamación que engloba toda la información pertinente a ella, a los cuales nos referiremos como “documentos finales”. Con esto, se concluye la fase de transformación y se procede al último procedimiento, conocido como “carga” (Load), que consiste en el almacenamiento estos documentos finales en la base de datos. Dejando todo el proceso de “Transform” como se muestra en la Figura 3.8.

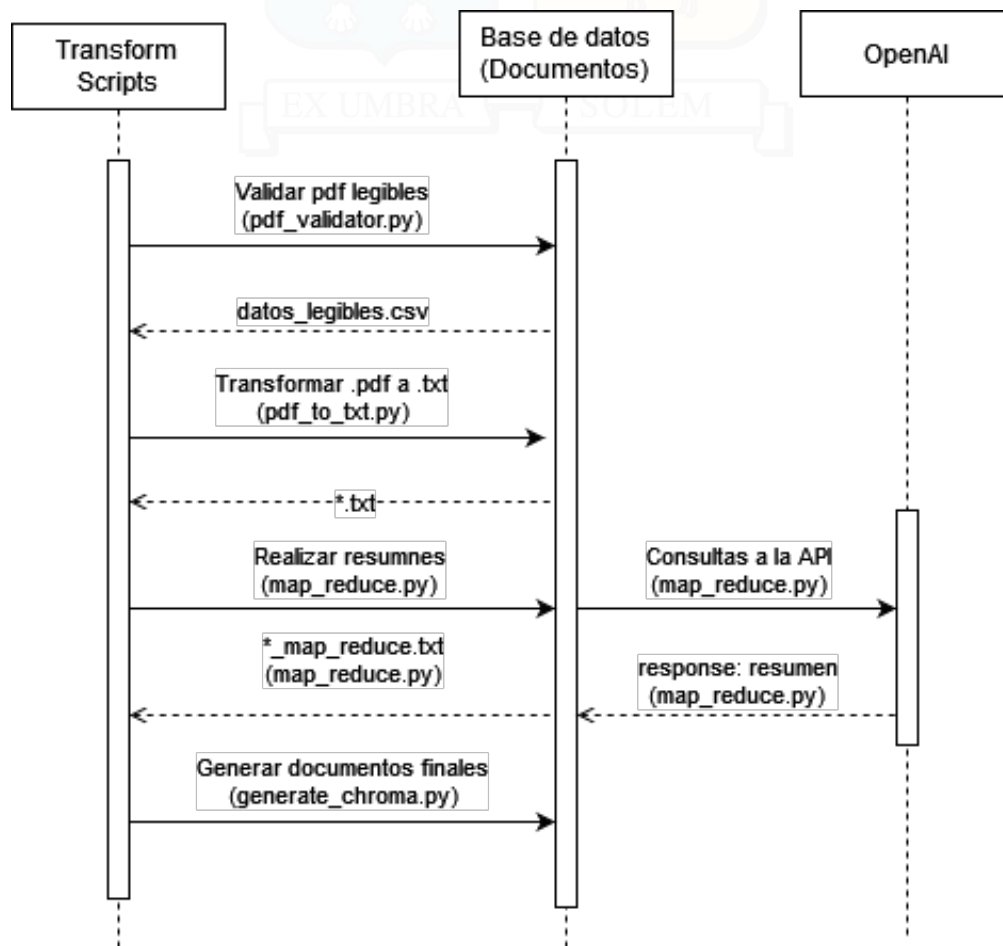


Figura 3.8: Diagrama de secuencia para el proceso de transformación (Transform) de datos

(Fuente: Elaboración propia)

3.1.3. Load

Al culminar el proceso de Extracción, Transformación y Carga (ETL), el siguiente y último paso a llevar a cabo la fase de carga, también conocida como “load” en inglés, en la cual se incorporan todos los documentos previamente descargados y transformados en una base de datos. Para este proyecto, en el cual se utiliza Langchain, resulta de vital importancia fragmentar los documentos en secciones más pequeñas con el propósito de que el Chatbot tenga la facilidad de consultar dichos fragmentos en un futuro, por lo que se deben dividir en fragmentos más pequeños todos los documentos a los que llamaremos chunk.

Esta necesidad de dividir estos documentos, surge debido a que los documentos deben ser sometidos a un proceso de Embeddings, proceso que previamente se explicó en el estado del arte de esta tesis, antes de ser introducidos en la base de datos. Esto se debe principalmente, a que las funciones de Embeddings tienen un límite en la extensión de grupos de caracteres, conocidos como “tokens”, que pueden ser procesados a la vez. En el contexto del modelo de Embeddings que funciona para el modelo de LLM “gpt-4” que utilizamos en este proyecto, se usa la función “text-embedding-ada-002” que cuenta con un límite que se establece en 8191 tokens [41], lo que constituye la longitud máxima de los fragmentos, o chunks, que formaremos en el proceso de “load”.

Por lo tanto, cuando se trabaja con documentos extensos, es indispensable dividirlos en fragmentos más pequeños antes de proceder con su incorporación, debido a las limitaciones de entrada en la ventana de contexto de los LLM, pues en este proyecto estos fragmentos después serán enviados como contexto y este contexto no puede exceder dicha ventana. Según la información proporcionada en el Blog de OpenAI, los embeddings son “representaciones numéricas de conceptos convertidos en secuencias numéricas, lo que facilita que las computadoras comprendan las relaciones entre los conceptos” [41]. En términos sencillos, los embeddings son representaciones vectoriales de texto que permiten su comprensión por parte de los Modelo de Grandes de Lenguaje (LLM). Dado que los LLM son redes neuronales, el proceso de Embedding resulta esencial para traducir el texto en números, que es el formato comprensible para esta red neuronal y permitir que el Chatbot funcione correctamente.

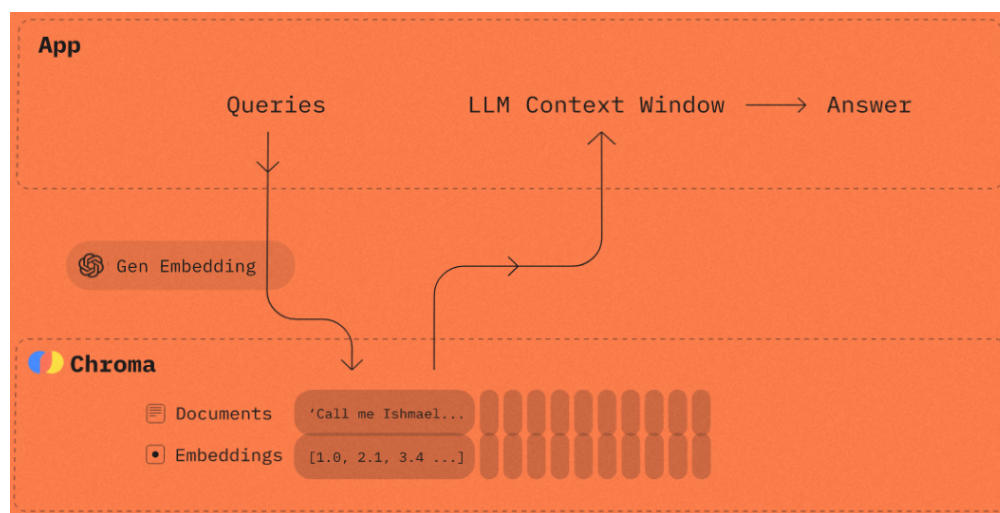


Figura 3.9: Diagrama de funcionamiento de ChromaDB

(Fuente: www.trychroma.com [14])

Langchain internamente, para llevar a cabo el proceso de carga de datos a la base de datos, toma todos los documentos finales proveniente del proceso de Transform en formato .txt y los divide en chunks que no superen la extensión máxima que puede procesar la función de embedding de OpenAI, para luego enviar uno por uno dichos chunk mediante llamadas a la API de OpenAI, de tal forma de obtener los vectores asociados a cada uno de los textos enviados. Los vectores resultantes de enviar los chunks de documentos a la función de embeddings, se almacenan posteriormente en una base de datos vectorial llamada ChromaDB. Esta base de datos vectorial, también llamada vecto-store en inglés, ha sido diseñada para ser compacta, escalable y eficiente, con el propósito de almacenar y recuperar vectores de manera efectiva. ChromaDB genera índices que permiten una recuperación rápida y eficiente de los embeddings en función de las consultas realizadas por los usuarios [52]. Con ello, esta base de datos guarda los vectores asociados al chunk de un documento, el texto del chunk del documento y su metadata asociada, tal como se muestra en la Figura 3.9.

Con este último procedimiento se concluye el proceso de ETL para los datos del tribunal ambiental, procediendo de esa manera a la programación e implementación del Chatbot que usará estos datos para entregar respuesta sobre estos datos con el uso de lenguaje natural y con ellos dejando el proceso de “load” como se muestra en la Figura 3.10.

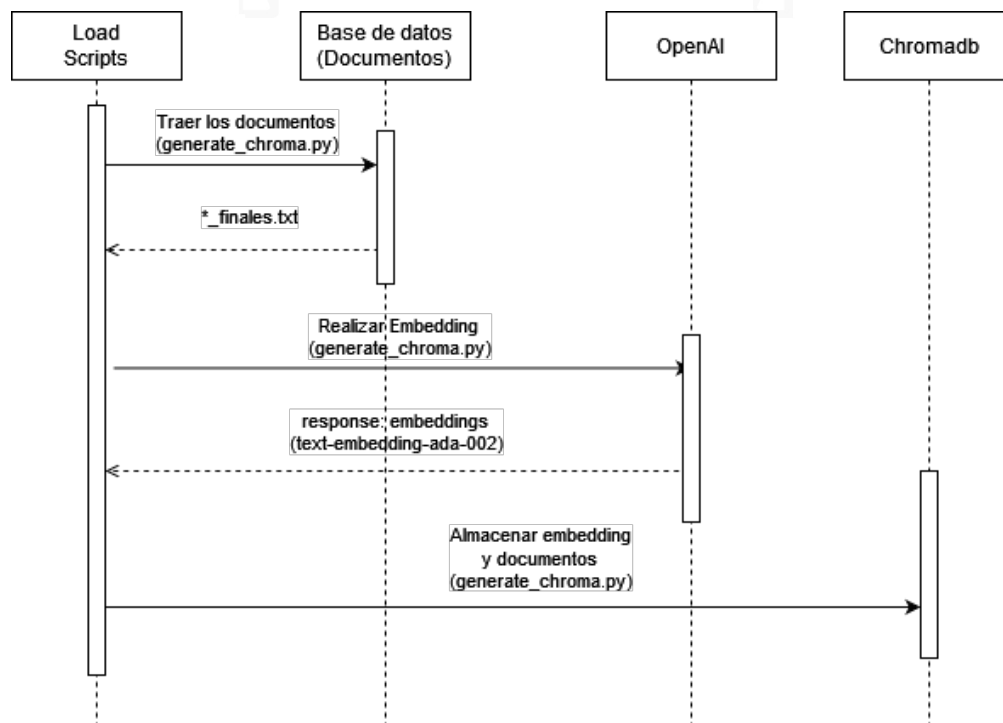


Figura 3.10: Diagrama de secuencia para el proceso de carga (Load) de datos

(Fuente: Elaboración propia)

3.2. Chatbot

Para la creación del Chatbot que utilizará los datos generados por el proceso de ETL anteriormene descrito, se utilizó una estructura de desarrollo formada en su totalidad por el lenguaje de programación Python [20], cosa que cada elemento fuera escrito en dicho lenguaje. La elección de este lenguaje de programación se debió a la experiencia por parte del equipo en Python, facilitando tanto la creación del frontend como del backend de la aplicación.

Para la sección de la interacción del usuario (frontend), se empleó el uso Python junto con el framework Flask [57] para la presentación de contenido en pantalla, incluyendo tanto la estructura de HTML como las hojas de estilo CSS. Además, se aprovechó y potenció el uso de Flask con el framework Bootstrap para agilizar el proceso de maquetación, ahorrando tiempo en el proceso de creación del Chatbot.

En cuanto a la creación del backend, se desarrolló una API con el objetivo de facilitar la interacción entre el frontend y el backend, aumentando así la velocidad de desarrollo a la vez que con ellos se aislaban posibles errores al desarrollar. Para este propósito, se utilizó FastAPI [55], un framework que permite el rápido desarrollo de APIs y que ofrece la ventaja de contar con Swagger [53] para la prueba y generación automática de documentación de la API.

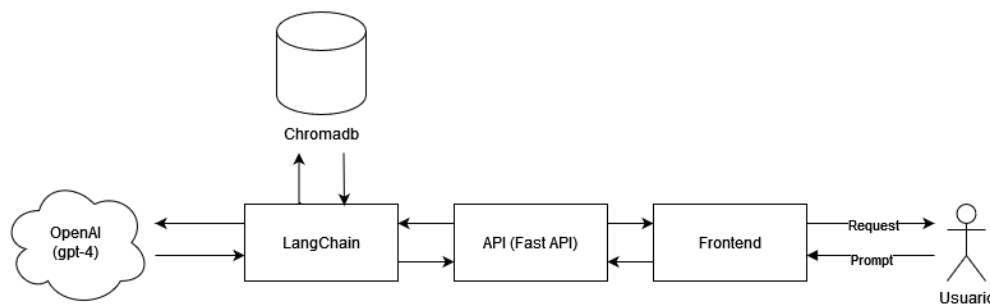


Figura 3.11: Diagrama de funcionamiento del Chatbot

(Fuente: Elaboración propia)

Para el funcionamiento interno de la aplicación, de la misma manera que se utilizó en el proceso de ETL, que este también funciona en base al framework Langchain, esto debido a que este framework al ser diseñado para el uso de LLMs y tareas asociadas a ellos, no solo funciona para el tratamiento de datos, sino que también para la interacción de los prompt con dichos LLM, dando también la posibilidad de integrar nuevas funciones como interacción con bases de datos, uso de código, etc [52]. Provocando que las aplicaciones realizadas con este framework se les llame LLM con “super poderes” gracias al abanico de funciones extras que aportan a los LLM.

Para la base de datos se utilizó ChromaDB, tal como se uso en el proceso de “load”. ChromaDB es una base de datos vectorial sin esquemas diseñada específicamente para su uso en aplicaciones de inteligencia artificial. Es liviano y muy potente, lo que permite el almacenamiento, la recuperación y la gestión eficiente de datos vectoriales (Embeddings), lo cual es esencial para las aplicaciones de chat de documentos basadas en LangChain y OpenAI [52]. Para que finalmente estas trabajaran en conjunto con los modelos de OpenAI, específicamente el modelo gpt-4 que actualmente es el más potente del mercado y que se utilizó para este Chatbot.

La función principal de este chatbot en específico es generar respuestas utilizando la técnica “Retriever-Augmented Generation” (RAG), que implica proporcionar contexto adicional en el prompt enviado a OpenAI. Mientras que una consulta simple suele constar de una instrucciones y una consulta, tal como se muestra en la Figura 3.12, en el uso RAG se agrega contexto dentro del prompt con el propósito de reducir la probabilidad de alucinaciones y mejorar la calidad de las respuestas.

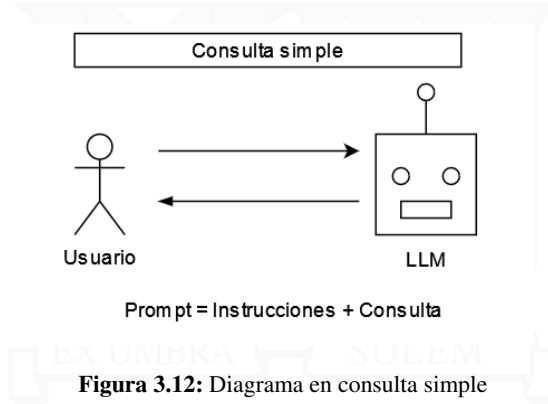


Figura 3.12: Diagrama en consulta simple

(Fuente: Elavoración propia)

El uso de “Retriever-Augmented Generation” (RAG) consiste en una consulta simple que al preguntar a un LLM se le hace entrega de las instrucciones y una consulta, pero que junto ello se agrega un contexto de este, cosa de que el modelo largo de lenguaje tenga menor capacidad de alucinar y generar una mejor respuesta, tal como se muestra en la Figura 3.13. Podemos decir que “Retriever-Augmented Generatio” (RAG) se refiere a un modelo de generación de lenguaje que se mejora por la capacidad de recuperar información de una base de datos, como un índice de vectores que representan los documentos del buscador ambiental, además de la memoria que ya posee. Este enfoque se utiliza para mejorar la generación de respuestas en tareas de procesamiento de lenguaje natural (NLP) que requieren conocimiento intensivo [32].

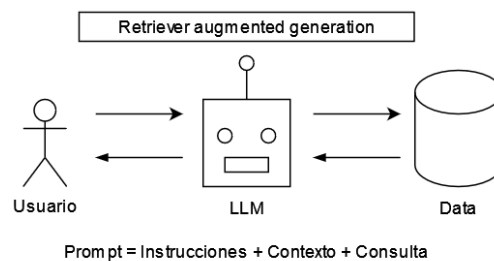


Figura 3.13: Diagrama en consulta mediante Retriver augmented generation

(Fuente: Elavoración propia)

Finalmente, este chatbot envía el prompt con las instrucciones y el contexto obtenido mediante el proceso de RAG, realizado con Langchain al LLM, siendo en este caso gpt-4 de OpenAI, luego de pasar por una función de Embedding, cosa de que pueda ser leído por el modelo. Por lo que se realiza un request con la información a la API de OpenAI, para que se obtenga el output con la respuesta, que postulamente será procesada y mostrada por el frontend.

3.3. Funcionamiento básico

Esta tesis se centra; tanto en el riesgo asociado al desarrollado y uso, como en el desarrollo de un Chatbot. Este chatbot funciona en base a la una arquitectura de RAG, Retriever-Augmented Generation en ingles, por lo que esta consiste en la recuperación de contexto, el cual es enviado junto con el prompt al LLM, por lo que se procede a dar un ejemplo del funcionamiento básico del chatbot. El proceso comienza con la obtención de un prompt específico del usuario por medio de la interfaz del frontend, como ejemplo usaremos el prompt: “Dame un resumen del documento Dominga” dentro de la barra de búsqueda del frontend, tal como se muestra en la Figura 3.14. Este prompt actúa como entrada inicial para el sistema de recuperación de información.

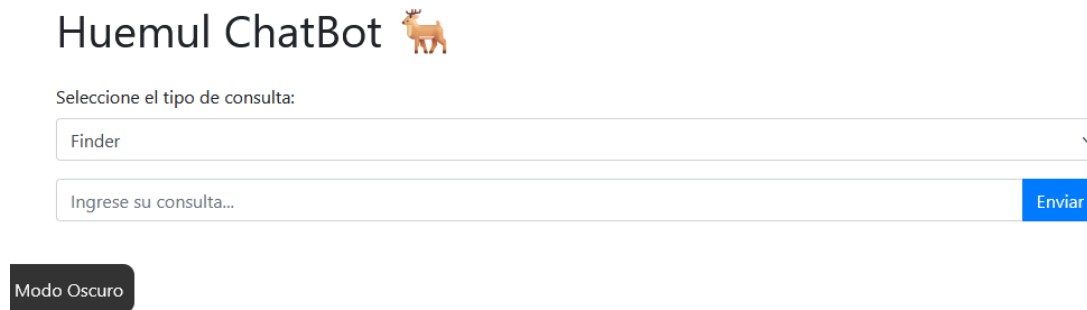


Figura 3.14: Screenshot del Chatbot desarrollado

(Fuente: Elaboración propia)

El prompt se procesa mediante una función de Embedding, empleando para ello OpenAI, siendo esta la función 'text-embedding-ada-002'. Esta función puede manejar hasta un máximo de 8191 tokens y produce un vector de 1536 dimensiones en forma de lista [41]. Para determinar la similitud entre el vector del prompt y los vectores correspondientes a los documentos almacenados, se utiliza la función de similitud coseno presente en la Ecuación 3.1. Esta mide el coseno del ángulo entre dos vectores, siendo estos A y B respectivamente, y este proporciona un valor que refleja su proximidad semántica entre el vector del prompt y los vectores de todos los documentos almacenados en la base de datos ChromaDB.

$$\text{similitud_coseno}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \quad (3.1)$$

El proceso de embedding como se mencionó anteriormente, lo que realiza en otras palabras es la conversión de un texto a un vector, esto debido a que los modelos de LLM al ser basados en redes neuronales necesitan una presentación gráfica de estos textos a un formato el cual pueda ser legible por el modelo, por lo que se convierten en números como muestra la Figura 3.15 a continuación.

```
prompt['embedding']
['[-0.019810765981674194, -0.00047473114682361484, -0.0085856292951107, 0.010154531199965954, -0.00646967125928783, -0.00527927977964282, -0.01359549164
7720337, 0.01686137355864048, -0.03711656853556633, -0.021628884598612785, 0.03531191498041153, 0.023258458822965622, -0.002772631822153926, 0.0187468305
2301407, -0.013804239220917225, 0.005989693105220795, 0.005972858518362045, -0.019177790731191635, 0.021655820310115814, -0.0015723364194855094, -0.00817
4803107976913, -0.011225203052163124, -0.025615280494089395, 0.006504826713353955, 0.0033820385579819785, 0.005999793764203787, 0.016349606215953827, -
0.0202013254166404, 0.01379077130516571, -0.002787782818628414, 0.019514480605721474, 0.001635465654545323, 0.0020958874374628807, -0.03881347924470901
5, -0.00596949178725481, -0.015366475097835064, -0.022342665120959282, -0.027554607018828392, 0.01574356676974297, -0.01978380270171165, -0.018288932740
68324, 0.0034746278543025255, 0.007959322072555556, -0.02777008985562325, 0.011918782256543636, -0.0007739632856100798, 0.010720170103013515, -0.005740
543361753225, -0.020470676343659782, 0.026921631768345833, -0.011595560237765312, 0.011407014913856983, -0.03399209678173065, -0.032079704105854034, 0.00
3628716669231653, 0.0047069089487195015, -0.02298910729587078, 0.009871713817119598, 0.002451094100251794, -0.0221271850168705, -0.005639536771923304, -
0.0199858434983101, -0.01842388379573822, 0.00013189818974844174, -0.007427353877574205, -0.01132943104207516, -0.00836966849118471, 0.009130998514592
640, -0.0005453575970456, 0.025642216209596294, 0.0273529383916855, 0.023447005078196526, 0.005740843361753225, -0.017836451026797295, 0.02393183670938
015, -0.000521025832315493, -0.0184505432844162, 0.005208575166761875, 0.015528805641562939, -0.0245513431727862, 0.01268643239251041, -0.0093965603772
99786, -0.00600984423186779, -0.00076137461881396, 0.035796746611595154, 0.033561136573553085, -0.00537691917270422, 0.01101645641028881, -0.017979178
577661514, -0.03100229986011982, -0.013817706145346165, 0.021480742841959, 0.01684790477156639, 0.018504414707422256, -0.03469241037964821, 0.02690816484
3916893, -0.01688830748740936, -0.0026581576094031334, 0.010262273252018345, -0.045412577688694, 0.005191740579903126, 0.0024645617231726646, -0.03132552
281022072, -0.012208334170281887, -0.0140277274967957, -0.026463735848665237, 0.021388481196268817, 0.0184094594570348167, -0.0004202296549919927, 0.00
0672114895880275, -0.015272201970219612, 0.01783103682100773, -0.02243639317989731, -0.023743290454149246, -0.004780980292707682, 0.0015378259122371674,
0.010787587519125938, -0.001114439801312983, -0.000709571526358438, -0.0535843190550804, 0.006551963277161121, 0.017332736402750015, 0.0215076769088168
8, -0.04988380521535734, 0.007346548605710268, -0.01684821048178673, -0.009400350041687489, -0.022840965539216995, -0.002363555133342743, 0.01053162384
7782612, 0.0112079543620348, 0.00666643725708127, -0.006151303648948669, -0.01769426272779054, -0.004881986882537693, 0.03401983063058853, -0.032564535
7370766, 0.02795863337814008, -0.03048443029575891, -0.04409275949001312, 0.019468649182715416, 0.0071176001800683, -0.01753474958240906, 0.0020117152
01669907, 0.03434225171804428, -0.0046631391160190105, 0.00702323275182545185, -0.007992991246283054, 0.008771783338487148, 0.01656508632004261, 0.022840
96539216995, -0.02838955448970795, -0.002105988096445799, 0.012672964483499527, -0.0054139550775289536, 0.00636884877681732, 0.014868175610899925, -0.023561330808337, -
001597588133583117, 0.00275579746812582, -0.0010403682244941592, 0.03897588978843689, -0.020336000248789787, 0.01564929261803627, 0.0223561330808337, -
```

Figura 3.15: Representación de un prompt luego de pasar por la función de embedding 'text-embedding-ada-002' en un archivo de Jupyter Notebook
(Fuente: Elaboración propia)

Luego del proceso anterior, el vector multinacional que representa al prompt debe ser comparado individualmente con cada uno de los vectores de embedding generados durante el proceso de ETL y que han sido previamente almacenados en la base de datos vectorial Chroma. Es crucial asegurarse de que tanto el vector asociado al prompt como los vectores asociados a los documentos almacenados hayan sido generados utilizando la misma función de embedding. Esta uniformidad es necesaria porque garantiza que ambos ocupen el mismo espacio vectorial y posean la misma dimensionalidad. Si los vectores no son del mismo tamaño, su comparación resulta inviable. Además, al haber sido generados bajo los mismos parámetros de la función de embedding, la evaluación de su semejanza se vuelve metodológicamente sólida. Como ejemplo podría darse una función de embedding que fue realizada solo con texto en inglés y una con varios idiomas, por lo que ambas tienen un carácter distintos y no sería útil comprarlas entre ellas.

Una vez calculada la similitud coseno entre el vector asociado al prompt y los vectores que presentan los documentos que se encontraban en la base de datos, se procede a elaborar un ranking de los documentos más relevantes, según la similitud obtenida del cálculo de similitud de cosenos. Esto se realiza seleccionando los "n" documentos con los valores de similitud más altos, siendo en el ejemplo proporcionado un total de 3 presente en la Figura 3.16.

```
df.sort_values("similarity", ascending=False).head(3)
```

	rol	nombreProyecto	page	page_content	embedding	prompt	color	similarity
49	R-1-2017	Dominga	0.0	El caso Dominga\n\nRol: "R-1-2017", Rol: "R-1-...	[-0.003615596564486623, -0.0018946133786812425...	R-1-2017	1	0.791294
40	R-2-2014	Loteo Riberas de la Dehesa	0.0	El caso Predio Tres Bocas Valdivia\n\nRol: "R-...	[0.0069757443852722645, -0.005744528956711292...		0	0.772962
36	R-283-2021	PTAS Gomero	0.0	El caso PTAS Gomero\n\nRol: "R-283-2021", Rol:...	[0.008394405245780945, -0.0023615628015249968...		0	0.772533

Figura 3.16: Screenshot de un jupyter notebook representando el proceso interno de selección

(Fuente: Elaboración propia)

Todo este proceso es orquestado internamente por Langchain, que consulta estos “n” documentos (siendo en este caso 3, aunque podrían ser más dependiendo de la preferencia y objetivo), a la base de datos vectorial ChormaDB y son enviados como contexto dentro del prompt a OpenAI mediante el uso de su API, junto a la key de autorización, para obtener el resultado del modelo y con ello una respuesta a la consulta en lenguaje natural.

Huemul ChatBot

Seleccione el tipo de consulta:

Finder

Dame un resumen del documento R-1-2017

Enviar

El documento R-1-2017 es un expediente relacionado con el proyecto minero-portuario Dominga en Chile. El caso involucra reclamaciones judiciales y recursos de casación presentados por diferentes partes. El Tribunal Ambiental ha revisado y analizado los antecedentes del proyecto, así como los argumentos técnicos utilizados para su rechazo. Se han discutido diversos aspectos, como la evaluación ambiental, la suficiencia de la información, las medidas de compensación, el valor compartido con las comunidades y los impactos ambientales. El Tribunal ha acogido parcialmente algunas alegaciones y ha desestimado otras, emitiendo una sentencia que ordena retrotraer el procedimiento de evaluación ambiental a una etapa posterior y realizar una nueva votación ajustada a derecho. El caso continúa en proceso y se espera una sentencia definitiva.

Figura 3.17: Screenshot del funcionamiento del Chatbot

(Fuente: Elaboración propia)

Finalmente, se obtiene el resultado enviado por parte del modelo, el cual llega en formato json tal como sale en el Apéndice A, este procesado por el frontend para que sea consumida de forma más amigable, de la forma que se observa en la Figura 3.17. Con el visto bueno del funcionamiento de la aplicación da fin al proceso de creación del Chatbot de principio a fin.

4 | Resultados

Esta tesis, al presentar una solución a un problema basado en el lenguaje natural (NLP), enfrenta el desafío de cuantificar los resultados entregados por el chatbot en respuesta a nuestras consultas. Debido a esta complicación, se propone utilizar una escala para evaluar dichos resultados en los ejemplos de la siguiente manera:

1. **Satisfactorio:** La respuesta es acorde a lo que menciona el prompt.
2. **Medianamente Satisfactoria:** La respuesta es adecuada, pero incluye errores que, si bien no invalidan la respuesta, podrían generar confusión.
3. **Insatisfactorio:** La respuesta no es pertinente o incluye información inventada.

Dada esta metodología para evaluar los casos, se presentarán ejemplos de prompts que un usuario podría emplear habitualmente en el uso de la aplicación. Se analizarán tanto el contexto enviado por el chatbot a OpenAI, como son resultados obtenidos y se realizará un contraste con la información de la que ya disponemos del período de desarrollo.

4.1. Caso 1: Consultar por resumen

Tabla 4.1: Primer ejemplo de resultado usando un prompt

Prompt	Respuesta
Dame un resumen del caso Dominga	El caso Dominga se refiere a una reclamación presentada por Andes Iron SpA contra la Resolución Exenta N° 1146/2017 del Servicio de Evaluación Ambiental y el Comité de Ministros. El reclamo se basa en el rechazo del Estudio de Impacto Ambiental del proyecto Dominga y la acogida parcial de los recursos de reclamación presentados en el marco de la participación ciudadana. El Primer Tribunal Ambiental de la República de Chile admitió la reclamación y convocó a una audiencia de percepción documental. Además, se presentaron informes de Amicus Curiae y se aceptó la participación de terceros coadyuvantes e independientes. El proceso judicial continúa y se espera una sentencia definitiva. En una sentencia anterior, el tribunal rechazó la excepción de extemporaneidad interpuesta por un tercero independiente y acogió la reclamación presentada por Andes Iron SpA. Sin embargo, la Corte Suprema anuló la sentencia y remitió el caso nuevamente al Primer Tribunal Ambiental para que se pronuncie sobre el fondo del asunto. El caso involucra diversos hechos controvertidos, como la evaluación del proyecto, la suficiencia de la información, los impactos ambientales, las medidas de compensación y la participación ciudadana.

Una de las consultas más evidentes podría ser la solicitud de un resumen de alguna de las reclamaciones mandadas al Tribunal; por lo tanto, es esencial evaluar el desempeño de este proceso con dicho prompt.

Como se puede observar en la respuesta proporcionada por el Chatbot, esta parece ser aceptable y lógica. Sin embargo, es necesario comprender ciertos aspectos. En primer lugar, enviar el prompt con el nombre del proyecto en lugar del código mejora la calidad de búsqueda, ya que los vectores de embedding reconocen más eficientemente las diferencias semánticas, proporcionando un contexto más adecuado. Además, es importante considerar que el envío de un exceso de contexto puede generar “alucinaciones”, resultando en resúmenes que no sean completamente precisos o que contengan información falsa generada por el LLM. El contexto enviado a OpenAI por el chatbot en el ejemplo y que contiene más contexto del necesario se encuentra en el Apéndice A. A pesar de los puntos anteriores, podemos decir que la respuesta entregada por el chatbot es **Satisfactoria**.

4.2. Caso 2: Consulta por listado

Tabla 4.2: Segundo ejemplo de un punteo o listado con todas la reclamaciones relacionadas con Humedales

Prompt	Respuesta
Dame un punteo o listado con todas la reclamaciones relacionadas con Humedales	A continuación se presenta un listado con todas las reclamaciones relacionadas con Humedales: 1. Caso Humedal La Marina: - Rol: “R-25-2021” - Fecha: 31-03-2022 - Reclamante: Inversiones F.K Limitada - Materia: Declaración humedal urbano - Resolución: Rechazada por el Third Environmental Court 2. Caso Humedal Urbano “Estero Agua Salada”: - Rol: “R-354-2022” - Fecha: 19-07-2023 - Reclamante: Consorcio Punta Puyai S.A. - Materia: Declaración humedal urbano - Resolución: Rechazada por el Second Environmental Court 3. Caso Humedal Quilicura: - Rol: “R-297-2021” - Fecha: 24-10-2022 - Reclamante: Inversiones Butamalal S.A. y otros - Materia: Declaración humedal urbano - Resolución: Acogida parcialmente por el Second Environmental Court 4. Caso Humedal El Bosque-Miraflores-Las Multas-Guacamayo: - Rol: “R-37-2021” - Fecha: 10-03-2023 - Reclamante: Nicolás Reichert Haverbeck - Materia: Declaración humedal urbano - Resolución: Acogida por el Third Environmental Court Espero que esta información sea de utilidad.

Otra consulta común es preguntar por listados a partir de la información del tribunal ambiental, este tiempo de preguntas resulta tener una dificultad engañosa para el Chatbot. Tras analizar la respuesta, se identifican aspectos tanto positivos como negativos. Entre los positivos, destaca que el Modelo de Lenguaje de Gran Escala (LLM) comprende correctamente el concepto de listado, generando una lista de cuatro resultados pertinentes. No obstante, se observa una limitación significativa: la base de datos vectorial incluye más de 20 documentos relacionados con humedales, pero debido a la restricción de enviar únicamente entre 3 a 5 documentos como contexto por parte del chatbot, no todos estos documentos pueden ser procesados para generar una respuesta correcta. Esta limitación resulta en una respuesta que, si bien es veraz, no es completa. Por lo tanto, aunque la información sobre humedales es incompleta, la respuesta puede calificarse como **Medianamente Satisfactoria**, basándose en la precisión de los datos proporcionados y la capacidad del sistema para entender y organizar la información relevante en un formato de listado.

4.3. Caso 3: Comparación entre reclamaciones

Tabla 4.3: Tercer ejemplo comparación entre reclamaciones

Prompt	Respuesta
Compara la reclamación R-25-2021 con la reclamación R-1-2017	La reclamación R-25-2021 se refiere al caso del Humedal La Marina, mientras que la reclamación R-1-2017 se refiere al caso de Dominga. Ambas reclamaciones son presentadas por diferentes partes en contra del Ministerio del Medio Ambiente y se centran en la evaluación ambiental de los respectivos proyectos. En el caso del Humedal La Marina, la reclamante argumenta que la solicitud de reconocimiento del humedal no cuenta con el acuerdo del Concejo Municipal y que se violó el debido proceso por falta de emplazamiento. Además, cuestiona la forma de notificación de los actos administrativos y la falta de motivación en la resolución reclamada. También acusa que la resolución incurre en desviación de poder y solicita que se deje sin efecto. Por otro lado, en el caso de Dominga, la reclamación se basa en el rechazo del Estudio de Impacto Ambiental del proyecto minero-portuario. Los reclamantes argumentan que se aplicaron incorrectamente los criterios de evaluación y que se omitieron estudios relevantes. También cuestionan la forma de notificación de los actos administrativos y la falta de motivación en la resolución reclamada. En este caso, la reclamante solicita que se modifique la resolución para no afectar su derecho de propiedad. En resumen, ambas reclamaciones se centran en la evaluación ambiental de los proyectos y cuestionan aspectos como la falta de acuerdo del Concejo Municipal, la forma de notificación de los actos administrativos, la falta de motivación en las resoluciones y la posible desviación de poder. Sin embargo, cada reclamación se basa en circunstancias y argumentos específicos relacionados con los proyectos y los humedales involucrados.

La propuesta de utilizar un prompt que busca comparar la relación entre dos textos representa una aplicación práctica y avanzada de un Modelo de Lenguaje de Gran Escala (LLM), más allá de su función como simple buscador. Esta metodología permite al LLM demostrar su habilidad para no solo identificar, sino también para analizar y comparar de manera fluida y sencilla los contenidos de diferentes documentos. En el ejemplo proporcionado, el LLM logra distinguir y describir cada texto antes de proceder a una comparación efectiva.

Sin embargo, se presenta una limitación similar a la observada en el primer ejemplo. Al revisar los contextos enviados a OpenAI revela que documentos no relacionados fueron incluidos en dicho contexto, lo cual, aunque no generó información incorrecta en este caso, tiene el potencial de causar alucinaciones. A pesar de este inconveniente, la capacidad del LLM para manejar adecuadamente los documentos relevantes y su eficacia en la comparación de textos justifican calificar la respuesta como **Satisfactoria**.

4.4. Caso 4: Consultar por información falsa

Tabla 4.4: Cuarto ejemplo preguntando por un texto que no existía

Prompt	Respuesta
Dame un resumen de R-14-2023	El resumen de R-14-2023 es el siguiente: La Resolución R-14-2023 se refiere al proyecto denominado Centro de Cultivo de Salmónidos “Córdova 4”. Esta resolución establece las regulaciones y requisitos para la operación y funcionamiento de dicho centro de cultivo. Entre las principales disposiciones de la resolución se encuentran: 1. Autorización para la construcción y operación del Centro de Cultivo de Salmónidos Córdova 4. ^{en} la ubicación especificada. 2. Establecimiento de medidas de protección ambiental para garantizar la preservación de los recursos naturales y la biodiversidad en el área de influencia del centro de cultivo. 3. Requisitos técnicos y de seguridad que deben cumplir las instalaciones del centro de cultivo, incluyendo sistemas de tratamiento de aguas residuales, control de calidad del agua y medidas de prevención de escapes de peces. 4. Obligaciones y responsabilidades del titular del proyecto en cuanto a la gestión adecuada de los desechos generados por el centro de cultivo y la implementación de programas de monitoreo y seguimiento ambiental. 5. Procedimientos y plazos para la presentación de informes periódicos sobre el estado y desempeño del centro de cultivo, así como para la realización de inspecciones y auditorías por parte de las autoridades competentes. Es importante destacar que este resumen es una síntesis de la resolución y no reemplaza la lectura y comprensión completa del documento original. Se recomienda consultar la resolución R-14-2023 en su totalidad para obtener información detallada sobre todas las disposiciones y requisitos establecidos.

Con el objetivo de ilustrar un aspecto crucial, se realizó una consulta al LLM utilizando información inexistente, específicamente la reclamación R-14-2023 que se inventó. Como era de esperarse, al solicitar información sobre un tema falso, el LLM generó una respuesta igualmente falsa. Este fenómeno se debe a que el chatbot ejecuta un proceso de semejanza semántica entre el prompt y los documentos disponibles, mandando estos documentos como contexto a OpenAI, esto pasa porque el chatbot no tiene como discernir si es necesario el proceso o no.

Este experimento subraya una limitación significativa del LLM: su incapacidad para discernir la veracidad de los datos en la consulta inicial. Dado que el modelo se basa en la información proporcionada, cualquier inexactitud en el prompt puede llevar a respuestas erróneas. Esta característica del LLM plantea preocupaciones particulares en contextos donde la fiabilidad y precisión de la información son fundamentales. En consecuencia, se considera que la respuesta obtenida en este caso es completamente **Insatisfactoria**.

4.5. Reflexiones generales de los resultados

Los resultados de los casos analizados, junto con el anexo presente en esta tesis, evidencian claramente las limitaciones del sistema RAG usado en este chatbot. Se considera que la gestión de la cantidad de información es problemática. Por un lado, proporcionar más información puede resultar en un contexto innecesariamente amplio y con eso dar paso a alucinaciones, mientras que ofrecer información insuficiente conduce a respuestas incompletas.

Tras una cuidadosa reflexión sobre estos resultados y el mecanismo de entrega de información, se identifican posibles mejoras para optimizar el rendimiento del chatbot. Una de las mejoras podría ser implementar un método que descarte contextos que no alcancen un cierto grado de similitud, aunque esto requeriría pruebas para determinar el umbral adecuado que diferencie los casos relevantes de los irrelevantes.

Otra mejora potencial sería analizar las respuestas insatisfactorias del chatbot y agregar documentos específicos a la base de datos. Por ejemplo, tomando el caso 2, se podría crear un documento que agrupe todos los casos relacionados con humedales, y de manera similar, generar documentos para cada tipo de elemento relevante. Sin embargo, esta solución presenta complejidades, como la identificación de elementos relevantes y el riesgo de saturar la base de datos.

Dando fin a los resultados, aunque este chatbot demuestra un desempeño prometedor y eficiente en la mayoría de los casos, existe un amplio margen para su mejora y optimización, con el fin de obtener resultados más precisos y de mayor calidad.

5 | Evaluación de Riesgos

En el dinámico campo de la Inteligencia Artificial, particularmente en el ámbito de los Modelos de Grandes de Lenguaje (LLM), la evaluación de riesgos es un elemento crucial para el éxito y la viabilidad de cualquier proyecto. Esta sección se dedica a explorar los desafíos y consideraciones esenciales en el desarrollo y aplicación de proyectos basados en LLM, ofreciendo una perspectiva integral sobre cómo mitigar riesgos y optimizar el rendimiento de estos modelos.

Al abordar la Evaluación de Riesgos en proyectos que involucran LLM, es indispensable considerar varios factores críticos que pueden influir directamente en el resultado del proyecto. Estos factores pueden aparecer tanto en el período de creación del proyecto, como mientras aplicamos la tecnología. A continuación, se presentarán una serie de riesgos que hay que tener en consideración para cualquier proyecto que utilice como base los LLM.

5.1. Creación del Proyecto

5.1.1. No tener un análisis previo de que se busca lograr

La realización de un análisis previo, tanto de una aplicación como un proyecto, resulta crucial para el inicio de cualquier proyecto basado en datos, independiente de la rama en donde se esté aplicando, donde el NLP no es la excepción. La mayoría de los usuarios que tienen acceso a herramientas analíticas, realmente no entienden el funcionamiento interno de dichas herramientas [23], por consiguiente, es crucial que al empezar un proyecto se entienda a la perfección donde se quiere llegar, que es lo que realmente necesita y cuáles son los objetivos.

Además, es común que suceda, que en un conjunto de datos muy extenso, cualquier efecto que desee probar aparecerá como significativo [23], de ahí la importancia de un análisis previo de objetivos y problemáticas. Este proceso es tan importante, debido a que existen casos en donde un problema que puede resultar complejo en primera instancia, puede sugerir el uso de modelos muy complejos. Sin embargo, a veces el uso de un modelo más simple, se llegan a resultados mejores que con el uso de modelos complejos [21], siendo este tipo de escenarios, un ejemplo claro de que un análisis previo adecuado puede mejorar tanto los resultados como la experiencia de desarrollo al trabajar.

5.1.2. Calidad de los datos

Para cualquier tipo de trabajo, aplicación o estudio, la calidad de los datos es en extremo importante debido a que, la calidad de los datos es crítica para un sistema de Machine Learning, porque los datos deficientes podrían causar problemas graves como predicciones erróneas o baja precisión de clasificación [15]. Ahora si hablamos del uso de LLM, que forman parte del procesamiento del lenguaje natural, además del Deep learning, simplificamos los atributos de calidad de datos en los tres más importantes para el Deep Learning: la fidelidad, la variedad y la veracidad de un conjunto de datos [15].

Como ejemplo, un caso que sucedió dentro del proyecto que acompaña a este análisis de riesgo, se encontraron problemas en una cierta cantidad de reclamaciones tratadas en el proceso de ETL, esto debido principalmente a que el Tribunal ambiental no cuenta con un estándar para entregar las reclamaciones que luego aparecerán en el buscador ambiental, existían tanto pdf que eran legibles, como otros que eran solo fotocopias, lo que imposibilitaba la extracción de información por problemas de formatos. El problema de la calidad de los datos a su vez, se puede extrapolar a la calidad de la metadata que era anexa en la base de datos dentro del proyecto.

5.1.3. Sesgos

Los Modelos de Lenguaje Grande (LLM), al ser entrenados con una masiva cantidad de datos, pueden manifestar sesgos debido al origen y procedencia de los datos utilizados en su entrenamiento. Estos sesgos pueden dar lugar a desafíos en desarrollos cuando se aplican en contextos distintos, ya que las respuestas generadas por el modelo pueden no ser adecuadas ni ajustarse a la realidad de esos nuevos escenarios, llevando así a respuesta erróneas, inexactas e incluso a respuestas discriminadoras.

Dentro de los muchos problemas que estos sesgos pueden generar algunos ejemplos son: Estereotipos de género en elección de ocupaciones, inexactitud ante las ambigüedades de una frase, dificultad de entender dinámicas complejas de género, sesgos culturales, discrepancia en las explicaciones de los modelos, etc. [29] Esto sucede porque los modelos preentrenados con corpus generados por humanos contienen sesgos sociales hacia ciertos grupos demográficos, porque el humano es un ser lleno de sesgos y con ello también los textos que produce. Estos sesgos son preocupantes, debido a que pueden ser propagados o incluso amplificadas en las tareas que estos modelos realizan [22], pudiendo llegar a malas repuestas o en el peor de los casos a problemas legales.

Como ejemplo podemos citar lo dicho por Bill Gates en su entrevista “Can AI Save the World? Expert Insights with Bill Gates” en donde menciono que: “Los sesgos en los modelos de IA pueden llevar a diagnósticos incorrectos, como se vio en el ejemplo donde Chat GPT diagnosticó erróneamente la tuberculosis como gripe debido a las bajas tasas de tuberculosis en los EE. UU.” [37]. Con este ejemplo, es que podemos dimensionar el efecto real de estos sesgos y en lo correcto o incorrecto que puede llegar a ser una respuesta por parte de estos modelos.

5.1.4. Elección correcta del modelo

Actualmente la oferta de grandes modelos de lenguaje tanto para investigación (research) como desarrollo comercial es muy amplia, podemos encontrar desde los privados como: ChatGPT, PaLM, Bloom, etc. Como también modelos de código abierto y libres para su uso comercial como: Llama 2, OpenLLaMA, Falcon, Dolly, etc. [36] Cada uno con sus respectivas variantes, debido a que existen variantes del modelo siendo unos más potentes que otros, a pesar

de ser de la misma familia. Como por ejemplo tenemos a Llama 2 que se puede encontrar en versión de 7, 13 y 70 billones de parámetros [30].

La elección de un modelo correcto para trabajar es sumamente importante para el desarrollo, esto debido a que tanto la diversidad y calidad de los datos de preentrenamiento influyen sustancialmente en la capacidad del modelo de lenguaje para comprender y proporcionar respuestas precisas, el tamaño puede tener una gran influencia en el rendimiento, además cada modelo tiene su soporte lingüístico [34], lo que dependiendo del tipo de proyecto podría ser crucial y determinante para el éxito de este.

5.1.5. Costos Monetarios

Los costos relacionados con la puesta en producción, en caso de usar un modelo open source, o el uso de un modelo LLM, considerando el uso de un LLM privado, pueden aumentar de manera exponencial si es que no se planea con antelación el costo monetario de ellos o no se predice la demanda de manera adecuada [51]. Por lo tanto, es fundamental tener en cuenta los costos asociados al utilizar un servicio de tercero, así como el costo de operar un servidor local, que pueden incluir el consumo de energía eléctrica, como la compra de tarjetas gráficas (GPU) para el funcionamiento correcto y óptimo de los modelos. Por ejemplo, existen estudios en donde al realizar ajustes al modelo, también llamado fine tuning, donde el consumo de energía es comparable al de pequeñas ciudades y el dióxido de carbono emitido es equivalente a 500 veces la de un vuelo de ida y vuelta entre Nueva York y San Francisco [24], por lo cual es un factor que hay que tener en cuenta.

5.1.6. Funciones de Embedding

Como se mencionó anteriormente en esta tesis, las funciones de Embedding son específicas para cada gran modelo de lenguaje y no son intercambiables entre modelos. Esto se debe a que los embeddings son representaciones de alto nivel provenientes de los pesos y parámetros de cada modelo, por lo que están diseñadas para captar y almacenar las relaciones semánticas específicas de cada modelo [31]. Por lo que, si se desea usar un modelo grande de lenguaje, es necesario contar con su función de Embedding correspondiente, además de la capacidad computacional para hacer funcionar dicha función, porque de lo contrario el modelo no podrá obtener el vector necesario de entrada para generar la predicción y con ello no podrá generar una respuesta.

5.1.7. Conocimiento de Framework

Cuando se desarrolla una aplicación, es esencial comprender el funcionamiento interno de los frameworks o herramientas que se utilizarán a lo largo del proyecto. Esta comprensión no solo es valiosa para entender cómo funciona el proceso en su conjunto, sino que también es necesaria para tener un control de los costos asociados al proyecto en caso de llevarse a producción.

Por ejemplo, en el contexto del proyecto de esta tesis, es importante saber que el framework Langchain, suele realizar múltiples llamadas a los modelos [12]. Si no se tiene conocimiento de esta situación o no se cuantifican de manera adecuada la cantidad de llamadas y la extensión de estas, esto debido a que los modelos como los entregados por OpenAI tiene un costo por los tokens que se reciben y por los que se envían [45], esto puede dar lugar a problemas en la cuantificación de costos. Por lo tanto, la capacidad de comprender y medir con precisión el uso de recursos, como las

llamadas a los modelos, es esencial para gestionar eficazmente los costos y asegurar el éxito del proyecto para cuando este sea enviado a producción.

5.1.8. Volatilidad del Mercado

Considerando a la fecha actual en donde se está escribiendo esta sección de la tesis, el 06 de noviembre de 2023, la creación de Chatbots utilizando la metodología RAG se perfilaba como una de las tendencias más destacadas en el mercado, siendo posiblemente una de las aplicaciones más prometedoras que los LLM tendría hasta la fecha. No obstante, en este mismo día, se llevó a cabo la OpenAI DevDay Keynote, donde se anunciaron novedades significativas en la oferta de productos por parte de la empresa OpenAI, como por ejemplo la entrada en escena de los GPTs, que permite personalizar versiones de ChatGPT con instrucciones, conocimiento extra y cualquier otra combinación de habilidades [44]. Junto con ello a su vez, se introdujeron nuevos modelos al catálogo de su API como “gpt-4-turbo”, un playground de desarrollo más completo para el desarrollo con herramientas de la empresa, text-to-speech (TTS), entre otros [46].

Con este contexto, es comprensible que comprometerse con cualquier tecnología conlleva riesgos, especialmente en este período caracterizado por una volatilidad extrema y una inversión extremadamente agresiva en inteligencia artificial. La Inteligencia Artificial Generativa continúa evolucionando de manera acelerada, lo que la hace cada vez más disruptiva y eficiente. Por lo tanto, la investigación y la implementación de soluciones de inteligencia artificial centradas en asistentes o chatbots representan un compromiso de alto riesgo en este entorno en constante cambio a la espera de nuevos servicios y competidores.

5.2. Uso de la Aplicación

5.2.1. Alucinaciones

En un contexto de uso de grandes modelos de lenguaje, el término “alucinación” se refiere a la generación de textos o respuestas que exhiben corrección gramatical, fluidez y autenticidad, pero se desvían de las entradas de fuente proporcionadas (fidelidad) o no se alinean con la precisión factual (factualidad) [61]. Explicando lo anteriormente dicho en términos más simples, decimos que un LLM alucina cuando las respuestas del modelo son coherentes y cohesiva, pero que sin embargo presenta información errónea o una lógica errada, también de una manera más coloquial podríamos decir que el modelo entrega información inventada a las preguntas que se le hacen.

Dicho lo anterior, podemos decir que este es un gran factor de riesgo para el uso de una aplicación, sobre todo si de su respuesta depende la toma de decisiones sensibles, porque a pesar de estar usando una metodología RAG, que dificulta la posibilidad de generar alucinaciones, sigue estando la posibilidad de que estas sucedan lo que puede entregar una respuesta con información errónea o inventada. Si es que las respuestas entregadas por un LLM no se revisan con criterio, se podría llegar a cometer graves errores al hacer uso de información que es directamente falsa.

5.2.2. Entrega de contexto adecuado

Los LLM a menudo suelen presentar alucinaciones, por lo que es esencial trabajar en reducir la frecuencia de este fenómeno. Para lograr dicho cometido, la provisión de contexto dentro del prompt para guiar la respuesta no es simplemente un capricho, sino que resulta absolutamente indispensable para la exactitud de la respuesta. De hecho, la entrega de contexto adecuado dentro del prompt ha demostrado ser una medida altamente efectiva para reducir las alucinaciones, logrando una disminución de hasta un 99.88 por ciento en algunos casos [16].

Por consiguiente, la correcta entrega de contexto dentro del prompt desempeña un papel fundamental en la generación de respuestas precisas a las consultas. Esto sucede principalmente porque, ya sea que el contexto proporcionado sea correcto, incorrecto o incluso irrelevante, el modelo de lenguaje lo utilizará como base para generar sus respuestas.

En el proyecto realizado en esta tesis, la generación de respuestas se basa por completo en la entrega de contexto dentro del prompt usando la metodología RAG, lo que a veces puede dar lugar a la transmisión de más información de la necesaria en el prompt debido al funcionamiento del framework de Langchain. Esto puede llevar a situaciones en las que el modelo, influenciado por la información incorrecta o adicional proporcionada, genere respuestas que no reflejan una respuesta en su totalidad correcta.

5.2.3. Limitaciones de la similitud de cosenos

La similitud del coseno, siendo este el método más usado por modelos RAG para la extracción de contexto en una base de datos vectorial, como medida de similitud semántica entre los vectores multidimensionales generados por una función de embeddings, particularmente para palabras de alta frecuencia en tareas de procesamiento de lenguaje natural (NLP) como preguntas y respuestas (QA), recuperación de información (IR) y traducción automática (MT) presenta limitaciones en su uso, esto principalmente sucede debido a que la frecuencia de las palabras en los datos de

entrenamiento afecta la geometría representacional de los embeddings contextualizados, siendo las palabras de baja frecuencias más concentradas geométricamente [63].

Por lo tanto, este problema se extrapola a que, en el momento de querer recuperar contexto pertinente de la base de datos vectorial, cuando se realiza el proceso de semejanza semántica por parte del RAG entre el prompt y los vectores de la base de datos, este pueda recibir información no relacionada con el prompt, debido a que los vectores que se solicitan siempre son un número fijo y siempre se extrae esa información eligiendo los valores de similitud más altos, por lo que se puede enviar más como contexto del necesario y puede dar oportunidad a alucinaciones.

5.2.4. Uso de información privada

Actualmente las empresas que entregan servicios de LLM son muy herméticas con la manera en que entrenan sus modelos, por lo tanto, no sabemos en un 100 por ciento con qué información han sido entrenados, la cual podría no necesariamente ser solamente información pública. Además, como ejemplo, todas las consultas realizadas en ChatGPT van directamente a los servidores de OpenAI y servirán posteriormente para reentrenar a los modelos de la compañía para mejorar su calidad y performance.

Este tema se ha vuelto tan delicado, que se ha probado que los ataques de reconstrucción de datos son posibles, en ellos se ha propuesto un ataque de reconstrucción dirigido de caja negra donde el adversario conoce parte de un ejemplo de entrenamiento (es decir, un indicio de texto) e infiere el resto (por ejemplo, un número de tarjeta de crédito), con lo que la posibilidad de extraer datos de entrenamiento de los modelos, lo que representa un riesgo serio para la privacidad [7].

Ante la duda de con qué información los modelos son entrenados, actualmente suceden casos como OpenAI que está siendo demandada tanto por violar los derechos de autor [59] como por robo sistemático [26] de información, quienes demandan alegan que sus obras han sido usadas para entrenar a sus modelos de LLM sin permiso de ellos y sin pago de regalías. Por lo que, hasta que no tengamos total transparencia del proceso de entrenamiento de estos modelos, a pesar de que existen servicios donde la información entregada “supuestamente” no es utilizada para reentrenar los modelos de la empresa [43], es preferible ser cautelosos con la información que se entrega estos LLM, siempre y cuando estos no estén corriendo de manera local, ya que eso asegura la privacidad de las consultas.

5.2.5. Aprendizaje por Refuerzo con Retroalimentación Humana (RLHF)

Los modelos grandes de lenguaje suelen dar respuestas que a veces, tanto política como moralmente no son correctas, posiblemente debido a la información con sesgos que fue utilizada para su entrenamiento, por lo que las empresas tienen por objetivo alinear los valores humanos con los sistemas de aprendizaje automático y dirigir los algoritmos de aprendizaje hacia los objetivos e intereses de los humanos [64]. A este proceso se le llama Aprendizaje por Refuerzo con Retroalimentación Humana (RLHF). Esta intervención de los resultados arrojados por los LLM, puede llegar a ser un problema si mismo, si se busca realizar una aplicación que tenga por usuario uno con una cultura diferente al proveedor del modelo, este podría generar respuestas no satisfactorias, incluso en caso de que un usuario no pudiera expresarse de manera correcta, podría hacer que el modelo se confunda y entregue resultados que podrían ser incompletos o incorrectos.

6 | Conclusiones

Tal como dijo Arthur C. Clarke: “Cualquier tecnología suficientemente avanzada es indistinguible de la magia”. Nos encontramos en un período donde la Inteligencia Artificial esta alcanzado capacidades que cada vez nos sorprenden más y nos asustan por igual, donde posiblemente nos estamos desilusionando al observar que aún no se logran cosas que imaginamos que a estas alturas ya deberían ser posibles, pero a su vez nos maravillamos viendo cómo se consiguen cosas que ni siquiera nos pudimos plantear en un inicio que se lograrían.

El uso comercial de la inteligencia artificial en especial de lo LLM, como fue la cobertura de esta tesis, presenta una cantidad muy alta de riesgos, tanto de los que son fáciles de prevenir, como pueden ser los cuantitativos dentro de los que se encuentran prevenir costos excesivos en servidores, como los que a su vez son difíciles de prevenir como las alucinaciones. Es importante para que estas aplicaciones tengan un buen futuro, entender el funcionamiento de estas como la manera en que fueron creadas. Los sesgos que presentan estos modelos no dejan de ser un reflejo de lo que somos como humanidad y que le dimos de alimento a estos modelos para ser entrenados, siendo las respuestas sesgadas entregadas por los modelos un gran reflejo de cómo somos nosotros y la manera en la que actuamos, demostrando que al igual que estos modelos somos lo que consumimos.

La industria de la Inteligencia Artificial deja expuesto a absolutamente todos los trabajos desde ahora en adelante en mayor o en menor medida, por lo que hay que tener cautela en las decisiones que se toman, pues el riesgo al que nos exponemos es sumamente grande. La volatilidad del mercado posiblemente es y será el riesgo más grande por considerar para cualquier tipo de proyecto de este tipo, no fue hace mucho que los llamados Prompt engineer serían los profesionales más cotizados en el mercado, incluso mencionados así por el CEO en Nvidia [18]. Sin embargo, estos fueron ya rápidamente reemplazados por los mismos LLM que se supone tenían que domar, debido a la optimización [60] o el auto mejoramiento mediante generación de prompt producidos el mismo LLM [17], dejando de esa manera obsoleto un rol que hace menos de tres meses a la fecha de publicación de esta tesis sería uno de los roles más importantes en el uso de Inteligencia Artificial Generativa.

Finalmente, para cualquier tipo de proyecto sobre o con uso de Inteligencia artificial siempre lo más importante serán los datos y el criterio del científico de datos detrás de ellos, porque existirán datos y herramientas, pero sin un conocimiento de mercado al que se apunta, realizar cualquier tipo de acción es trabajar en la oscuridad porque sin criterio, trabajar con datos es un trabajo en vano y sin sentido.

7 | Recomendaciones

Queda propuesto para quien quiera continuar con esta tesis el generar resultados que puedan ser medidos de mejor manera y no tan cualitativamente como se realizó en esta tesis para estudiar la viabilidad de usar en producción este tipo de chatbot, también queda propuesto el solucionar el problema con la semejanza semántica usando proximidad de cosenos, cosa de no traer al contexto información que no era necesaria que luego será enviado al servidor. Finalmente, el uso de MLOps para la mantención y optimización de este tipo de chatbot, para que el uso continuo de la aplicación cada vez genere mejores respuestas.

Bibliografía

- [1] ¿Qué es la IA generativa y cuáles son sus aplicaciones? | Google Cloud. URL: <https://cloud.google.com/use-cases/generative-ai?hl=es>.
- [2] Eleni Adamopoulou y Lefteris Moussiades. “An overview of chatbot technology”. En: *IFIP international conference on artificial intelligence applications and innovations*. Springer. 2020, págs. 373-383.
- [3] Ian L. Alberts y col. “Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be?”. En: *European journal of nuclear medicine and molecular imaging* 50.6 (2023), págs. 1549-1552.
- [4] Syed Muhammad Fawad Ali y Robert Wrembel. “From conceptual design to performance optimization of ETL workflows: current state of research and open problems”. En: *VLDB Journal* 26 (6 dic. de 2017), págs. 777-801. ISSN: 0949877X. DOI: [10.1007/S00778-017-0477-2](https://doi.org/10.1007/S00778-017-0477-2). URL: <https://link.springer.com/article/10.1007/s00778-017-0477-2>.
- [5] Xavier Amatriain y col. “Transformer models: an introduction and catalog”. En: (feb. de 2023). URL: <https://arxiv.org/abs/2302.07730v3>.
- [6] Tribunal Ambiental. *Tribunal escuchó alegatos en 7 reclamaciones que buscan anular la aprobación ambiental del proyecto minero Blanco*. URL: <https://tribunalambiental.cl/audiencia-r-333-2022-acumula-6-rca-proyecto-blanco-atacama/>.
- [7] Borja Balle, Giovanni Cherubin y Jamie Hayes. “Reconstructing Training Data with Informed Adversaries”. En: *Proceedings - IEEE Symposium on Security and Privacy 2022-May* (2022), págs. 1138-1156. ISSN: 10816011. DOI: [10.1109/SP46214.2022.9833677](https://doi.org/10.1109/SP46214.2022.9833677).
- [8] Tom Brown y col. “Language models are few-shot learners”. En: *Advances in neural information processing systems* 33 (2020), págs. 1877-1901.
- [9] *Buscador Ambiental*. URL: <https://www.buscadorambiental.cl/buscador/#/>.
- [10] Yi Cao y Jia Zhai. “Bridging the gap—the impact of ChatGPT on financial research”. En: *Journal of Chinese Economic and Business Studies* 21 (2 2023), págs. 177-191. ISSN: 14765292. DOI: [10.1080/14765284.2023.2212434](https://doi.org/10.1080/14765284.2023.2212434). URL: <https://www.tandfonline.com/action/journalInformation?journalCode=rcea20>.
- [11] Armanda Cetrulo y Alessandro Nuvolari. “Industry 4.0: revolution or hype? Reassessing recent technological trends and their impact on labour”. En: *Journal of Industrial and Business Economics* 46 (3 sep. de 2019), págs. 391-402. ISSN: 19724977. DOI: [10.1007/S40812-019-00132-Y](https://doi.org/10.1007/S40812-019-00132-Y). URL: <https://link.springer.com/article/10.1007/s40812-019-00132-y>.
- [12] Harrison Chase. *Welcome to LangChain*. <https://langchain-doc.readthedocs.io/en/latest/index.html>. Accessed: 2023-11-06. 2022.
- [13] Abhimanyu Chopra, Abhinav Prashar y Chandresh Sain. “Natural language processing”. En: *International journal of technology enhancements and emerging engineering research* 1.4 (2013), págs. 131-134.
- [14] ChormaDB. *the AI-native open-source embedding database — trychroma.com*. <https://www.trychroma.com/>. [Accessed 25-11-2023].
- [15] Junhua Ding y col. “A case study of the augmentation and evaluation of training data for deep learning”. En: *Journal of Data and Information Quality* 11 (4 ago. de 2019). ISSN: 19361963. DOI: [10.1145/3317573](https://doi.org/10.1145/3317573). URL: <https://doi.org/10.1145/3317573>.

- [16] Philip Feldman, James R. Foulds y Shimei Pan. “Trapping LLM Hallucinations Using Tagged Context Prompts”. En: (jun. de 2023). URL: <https://arxiv.org/abs/2306.06085v1>.
- [17] Chrisantha Fernando y col. *Promptbreeder: Self-Referential Self-Improvement via Prompt Evolution*. 2023.
- [18] Yahoo Finance. *Nvidia’s CEO just gave a graduation speech about the future of work and said that A.I. won’t steal jobs but ‘someone who’s an expert with A.I. will’*. 2023. URL: <https://finance.yahoo.com/news/nvidia-ceo-just-gave-graduation-183507133.html>.
- [19] Diario Financiero. *PAPEL DIGITAL*. URL: <https://www.df.cl/empresas/energia/hidroaysen-sufre-reves-judicial-en-defensa-de-derechos-de-agua> (visitado 19-11-2023).
- [20] Python Software Foundation. *Welcome to Python.org* — *python.org*. <https://www.python.org/>. [Accessed 25-11-2023].
- [21] Edward J. GREGG y col. “Why less complexity produces better forecasts: an independent data evaluation of kelp habitat models”. En: *Ecography* 42 (3 mar. de 2019), págs. 428-443. ISSN: 16000587. doi: [10.1111/ECOG.03470](https://doi.org/10.1111/ECOG.03470).
- [22] Yue Guo, Yi Yang y Ahmed Abbasi. “Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts”. En: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. por Smaranda Muresan, Preslav Nakov y Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, mayo de 2022, págs. 1012-1023. doi: [10.18653/v1/2022.acl-long.72](https://doi.org/10.18653/v1/2022.acl-long.72). URL: <https://aclanthology.org/2022.acl-long.72>.
- [23] D HAND. “Intelligent data analysis: Issues and opportunities”. En: *Intelligent Data Analysis* 2 (1-4 ene. de 1998), págs. 67-79. ISSN: 1088-467X. doi: [10.1016/S1088-467X\(99\)80001-8](https://doi.org/10.1016/S1088-467X(99)80001-8).
- [24] Kai Huang y col. “Towards Green AI in Fine-tuning Large Language Models via Adaptive Backpropagation”. En: *arXiv preprint arXiv:2309.13192* (2023).
- [25] Whitney Hunt, Kendal Marshall y Ryan Perry. *Artificial Intelligence’s Role in Finance and How Financial Companies are Leveraging the Technology to Their Advantage*. Disponible en SSRN: <https://ssrn.com/abstract=3707908>. 2020.
- [26] Infobae. *Escritor de Juego de Tronos demanda a OpenAI por “robo sistemático”*. URL: <https://www.infobae.com/tecnologia/2023/09/23/escritor-de-juego-de-tronos-demanda-a-openai-por-robo-sistemático/>.
- [27] Mladan Jovanovic y Mark Campbell. “Generative artificial intelligence: Trends and prospects”. En: *Computer* 55.10 (2022), págs. 107-112.
- [28] Kaspersky. *What are bots? - definition and explanation*. Abr. de 2023. URL: <https://www.kaspersky.com/resource-center/definitions/what-are-bots>.
- [29] Hadas Kotek, Rikker Dockum y David Sun. “Gender Bias and Stereotypes in Large Language Models”. En: *Proceedings of The ACM Collective Intelligence Conference. CI ’23*. Delft, Netherlands: Association for Computing Machinery, 2023, págs. 12-24. ISBN: 9798400701139. doi: [10.1145/3582269.3615599](https://doi.org/10.1145/3582269.3615599). URL: <https://doi.org/10.1145/3582269.3615599>.
- [30] Lakera. *The List of 11 Most Popular Open Source LLMs of 2023 | Lakera – Protecting AI teams that disrupt the world*. URL: <https://www.lakera.ai/blog/open-source-llms>.
- [31] Microsoft Learn. *LLM AI Embeddings*. URL: <https://learn.microsoft.com/en-us/semantic-kernel/memories/embeddings>.
- [32] Patrick Lewis y col. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. En: *Advances in Neural Information Processing Systems* 33 (2020), págs. 9459-9474. URL: <https://github.com/huggingface/transformers/blob/master/>.
- [33] *Ley Chile - Ley 20600 - Biblioteca del Congreso Nacional*. URL: <https://www.bcn.cl/leychile/navegar?idNorma=1041361&idParte=9269911>.
- [34] Shreekanth Mandvikar. “Factors to Consider When Selecting a Large Language Model: A Comparative Analysis”. En: *International Journal of Intelligent Automation and Computing* 6 (3 ago. de 2023), págs. 37-40. URL: <https://research.tensorgate.org/index.php/IJIAC/article/view/53>.
- [35] Puranjay Savar Mattas. “ChatGPT: A Study of AI Language Processing and its Implications”. En: *International Journal of Research Publication and Reviews* 04 (02 2023), págs. 435-440. doi: [10.55248/GENGPI.2023.4218](https://doi.org/10.55248/GENGPI.2023.4218).

- [36] Meta. *Llama 2 - Meta AI*. URL: <https://ai.meta.com/llama/>.
- [37] Mrwhosetheboss. *Can AI really save the World? ft. Bill Gates - YouTube*. URL: <https://www.youtube.com/watch?v=l9m3IKG8i88&t=16s>.
- [38] N. K. Nagwani. "Summarizing large text collection using topic modeling and clustering based on MapReduce framework". En: *Journal of Big Data* 2 (1 dic. de 2015), págs. 1-18. ISSN: 21961115. DOI: [10.1186/s40537-015-0020-5](https://doi.org/10.1186/s40537-015-0020-5). URL: <https://link.springer.com/articles/10.1186/s40537-015-0020-5>. URL: <https://link.springer.com/article/10.1186/s40537-015-0020-5>.
- [39] Arvind Neelakantan y Lilian Weng. *Introducing text and code embeddings*. <https://openai.com/blog/introducing-text-and-code-embeddings>. Accessed: 2023-11-16. Ene. de 2022.
- [40] Arvind Neelakantan y col. "Text and Code Embeddings by Contrastive Pre-Training". En: *arXiv preprint arXiv:2201.10005* arXiv:2201.10005 (ene. de 2022). URL: <https://arxiv.org/abs/2201.10005>.
- [41] OpenAI. *Embeddings - OpenAI API*. URL: <https://platform.openai.com/docs/guides/embeddings/what-are-embeddings>.
- [42] OpenAI. *Introducing ChatGPT*. <https://openai.com/blog/chatgpt>. Accessed: 2023-11-16. Nov. de 2022.
- [43] OpenAI. *Introducing ChatGPT Enterprise*. URL: <https://openai.com/blog/introducing-chatgpt-enterprise>.
- [44] OpenAI. *Introducing GPTs*. URL: <https://openai.com/blog/introducing-gpts>.
- [45] OpenAI. *Models*. <https://platform.openai.com/docs/models/>. [Accessed 25-11-2023]. 2023.
- [46] OpenAI. *New models and developer products announced at DevDay*. URL: <https://openai.com/blog/new-models-and-developer-products-announced-at-devday>.
- [47] OpenAI. *What are Embeddings*. <https://platform.openai.com/docs/guides/embeddings/what-are-embeddings>. Accessed: 2023-11-16. Nov. de 2023.
- [48] Alec Radford y col. "Improving language understanding by generative pre-training". En: (2018).
- [49] Frank Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain." En: *Psychological review* 65.6 (1958), pág. 386.
- [50] Selenium. *Selenium - selenium.dev*. <https://www.selenium.dev/>. Acceso en 25-11-2023.
- [51] Jeremy Smart. *Cloud Pricing Model: How to Budget and Forecast Cloud Computing Cost*. URL: <https://primetr.com/insights/common-cloud-cost-models-how-to-budget-and-forecast-cloud-computing-cost/> (visitado 14-12-2023).
- [52] Mirela Șorecău y Emil Șorecău. "AN ALTERNATIVE APPLICATION TO CHATGPT THAT USES RELIABLE SOURCES TO ENHANCE THE LEARNING PROCESS". En: XXIX (2023), pág. 2023. DOI: [10.2478/kbo-2023-0084](https://doi.org/10.2478/kbo-2023-0084).
- [53] Swagger. *API Documentation & Design Tools for Teams | Swagger — swagger.io*. <https://swagger.io/>. [Accessed 25-11-2023].
- [54] La Tercera. *SEA pide a Tribunal Ambiental rechazar reclamación de Andes Iron por Dominga*. URL: <https://www.latercera.com/pulso-pm/noticia/sea-pide-a-tribunal-ambiental-rechazar-reclamacion-de-andes-iron-por-dominga-y-acusa-a-la-minera-de-permanente-victimizacion/ZJ6KKJLV6ZDJFM6YKGAUF6PP7E/>.
- [55] Tiangolo. *FastAPI — fastapi.tiangolo.com*. <https://fastapi.tiangolo.com/>. [Accessed 25-11-2023].
- [56] Ashish Vaswani y col. "Attention Is All You Need". En: (2023).
- [57] Werkzeug. *Welcome to Flask; Flask Documentation (3.0.x) — flask.palletsprojects.com*. <https://flask.palletsprojects.com/en/3.0.x/>. [Accessed 25-11-2023].
- [58] *What is script? | Definition from TechTarget — techtarget.com*. <https://www.techtarget.com/whatis/definition/script>. [Accessed 25-11-2023].
- [59] WIRED. *Comediante Sarah Silverman demanda a OpenAI y Meta por infringir derechos de autor*. URL: <https://es.wired.com/articulos/sarah-silverman-demanda-a-openai-y-meta-por-infringir-derechos-de-autor>.

- [60] Chengrun Yang y col. *Large Language Models as Optimizers*. 2023.
- [61] Hongbin Ye y col. "Cognitive Mirage: A Review of Hallucinations in Large Language Models". En: (sep. de 2023). URL: <https://arxiv.org/abs/2309.06794v1>.
- [62] J.D. Zamfirescu-Pereira y col. "Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts". En: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. <conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: [10.1145/3544548.3581388](https://doi.org/10.1145/3544548.3581388). URL: <https://doi.org/10.1145/3544548.3581388>.
- [63] Kaitlyn Zhou y col. "Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words". En: *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2* (mayo de 2022), págs. 401-423. ISSN: 0736587X. DOI: [10.18653/v1/2022.acl-short.45](https://arxiv.org/abs/2205.05092v1). URL: <https://arxiv.org/abs/2205.05092v1>.
- [64] Banghua Zhu, Jiantao Jiao y Michael I. Jordan. "Principled Reinforcement Learning with Human Feedback from Pairwise or K -wise Comparisons". En: (ene. de 2023). ISSN: 26403498. URL: <https://arxiv.org/abs/2301.11270v4>.

A | Contexto enviado por el Chatbot

Elaboración propia - Ejemplo de Contexto enviado por el Chatbot a OpenAI

```
{
  'query': 'Dame un resumen del caso Dominga',
  'result': 'El caso Dominga se refiere a una reclamación presentada por Andes Iron SpA
contra la Resolución Exenta N° 1146/2017 del Servicio de Evaluación Ambiental y el Comité de
Ministros. El reclamo se basa en el rechazo del Estudio de Impacto Ambiental del proyecto
Dominga y la acogida parcial de los recursos de reclamación presentados en el marco de la
participación ciudadana. El Primer Tribunal Ambiental de la República de Chile admitió la
reclamación y convocó a una audiencia de percepción documental. Además, se presentaron
informes de Amicus Curiae y se aceptó la participación de terceros coadyuvantes e
independientes. El proceso judicial continúa y se espera una sentencia definitiva. En una
sentencia anterior, el tribunal había anulado varias resoluciones relacionadas con el proyecto
y ordenado retrotraer el procedimiento de evaluación ambiental a una etapa posterior al
Informe Consolidado de Evaluación. Sin embargo, esta sentencia fue anulada por la Corte
Suprema y el caso fue remitido nuevamente al Primer Tribunal Ambiental para que se pronuncie
sobre el fondo del asunto.',
  'source_documents': [
    Document(
      page_content='El caso Dominga\n\nRol: "R-1-2017", Rol: "R-1-2017", Rol: "R-1-2017",
fecha 16-04-2021 de Andes Iron SpA con SEA cuyo proyecto corresponde a Dominga. Este caso
tiene como principal materia Evaluación Ambiental y su coponente ambiental es Agua\n\nThe
integration ministers are: \n1. Juan Opazo Lagos\n2. Jasna Pavlich Núñez\n3. Fabrizio Queirolo
Pellerano\n\nSiendo sus descriptores: Evaluación ambiental, término anticipado, suficiencia de
información, medidas de compensación, valor compartido, facultades del Comité de Ministros,
suficiencia de la información, línea base de medio marino, medida voluntaria, riesgos
ambientales, observación ciudadana, PAC, fauna marina, impacto de ruidos, rutas de navegación,
efectos sinérgicos, áreas de exclusión, trato desigual, biodiversidad marina, pinguino
Humboldt, barreras eólicas, amicus curiae. \n\nEl First Environmental Court Acoge la causa en
esta ocacion. Los hechos controvertidos son: Del supuesto carácter completo de los
antecedentes del proyecto y evaluación del mismo. En cuanto al término anticipado del
procedimiento de evaluación ambiental. En cuanto al criterio del propio Comité de Ministros
respecto de los aspectos relacionados con la actividad de transporte marítimo. De la supuesta
información para evaluar las rutas de navegación que habría cumplido con creces el estándar
fijado por el Comité de Ministros. En relación a la propuesta de la casa del Mar, luego
denominada centro de estudios científicos. En cuanto al "Valor Compartido" con las
comunidades. Revisión de los argumentos técnicos que fundaron el rechazo del proyecto. De la
supuesta insuficiencia de la línea de base del medio marino en las rutas de navegación. De la
supuesta utilización parcial de fragmentos contenidos en los informes de los Servicios, con
preguntas inductivas al efecto. Del supuesto tratamiento discriminatorio en lo relativo a la
línea de base. De la supuesta suficiencia de los análisis de efectos sinérgicos. De la
supuesta ausencia de desequilibrio entre impactos y medidas. De las Reclamaciones PAC. Del
supuesto impacto en el recurso hídrico luego del cierre del Sistema de Captación y Devolución
de aguas desde el Rajo Sur. De la supuesta claridad de la información para descartar impactos
adversos en el mar. De la supuesta insignificancia de impactos por aporte de material
particulado sedimentable ("MPS"). De la supuesta insuficiencia de la medida de compensación
para el impacto en el proceso de nidificación de pinguino de Humboldt. De la supuesta
ilegalidad en la posición sobre la evaluación del impacto de barreras eólicas en la avifauna.
De la supuesta suficiencia de las áreas de exclusión. De la supuesta evaluación del impacto
acústico de los buques en la fauna marina. Del supuesto plan de monitoreo de aves y cetáceos
que no podría supuestamente ser medida.\n\nY el resumen es: El Primer Tribunal Ambiental de la
República de Chile recibió una reclamación presentada por Andes Iron SpA contra la Resolución
Exenta N° 1146/2017 del Servicio de Evaluación Ambiental y el Comité de Ministros. La
reclamación se basa en el rechazo del Estudio de Impacto Ambiental del proyecto Dominga y la
acogida parcial de los recursos de reclamación presentados en el marco de la participación
ciudadana. El Tribunal admitió a trámite la reclamación y convocó a una audiencia de
percepción documental. Además, se presentaron informes de Amicus Curiae y se aceptó la
participación de terceros coadyuvantes e independientes. El proceso judicial continúa y se
espera una sentencia definitiva.\n\nEl Primer Tribunal Ambiental de Chile ha emitido una
sentencia en relación al proyecto minero-portuario Dominga. En la sentencia, se rechaza la
excepción de extemporaneidad interpuesta por un tercero independiente y se acoge la
reclamación presentada por Andes Iron SpA. Se anulan varias resoluciones relacionadas con el
proyecto y se ordena retrotraer el procedimiento de evaluación ambiental a una etapa posterior
al Informe Consolidado de Evaluación. La Corte Suprema ha anulado la sentencia y ha remitido
el caso nuevamente al Primer Tribunal Ambiental para que se pronuncie sobre el fondo del
asunto.\n\nEl documento revisado es una serie de antecedentes procesales de un caso relacionado
con el proyecto minero Dominga en Chile. El caso involucra reclamaciones judiciales y recursos
de casación presentados por diferentes partes. La Corte Suprema ha ordenado al Tribunal
Ambiental que revise y decida sobre el fondo del asunto, incluyendo la evaluación ambiental
del proyecto y la refutación de los antecedentes técnicos utilizados para rechazarlo. Los
principales hechos controvertidos incluyen el carácter completo de los antecedentes del
proyecto, la evaluación del mismo y los argumentos técnicos que fundamentaron su rechazo. El
```


documento también menciona la suspensión de la vista de la causa y la fijación de una nueva fecha para la audiencia de alegatos. El Tribunal Ambiental de Chile revisó una serie de alegaciones relacionadas con la evaluación ambiental del proyecto Dominga. En primer lugar, se discutió si el procedimiento de evaluación se terminó anticipadamente, concluyendo que no existió falta de información relevante. Además, se argumentó que el proyecto cumplió con todas las instancias formales de participación ciudadana. En segundo lugar, se analizó el criterio del Comité de Ministros respecto a la actividad de transporte marítimo del proyecto, concluyendo que el Comité consideró que la línea de base complementaria para las rutas de navegación no fue suficiente para evaluar los efectos del proyecto. Sin embargo, el Tribunal consideró que la línea de base del medio marino presentada por el titular del proyecto fue suficiente y adecuada. Por último, se discutió la propuesta de la Casa del Mar, concluyendo que el proyecto consideró compromisos voluntarios para cumplir con estándares más altos que los establecidos por la legislación ambiental. En resumen, el Tribunal acogió algunas alegaciones del reclamante y desestimó otras, concluyendo que el proyecto Dominga cumplió con los requisitos de evaluación ambiental. El Tribunal Ambiental ha revisado los argumentos técnicos que fundamentaron el rechazo del proyecto Dominga. En cuanto a la medida propuesta de crear un Centro de Estudios Científicos, el tribunal considera que es interesante e innovadora, pero sugiere que debería ser más amplia en su alcance y concepción, abarcando también el estudio de los ecosistemas terrestres y el medio humano en la comuna de La Higuera. Además, se propone que el Centro de Estudios lleve a cabo un programa de investigación y estudios de las especies de interés en la zona, así como un protocolo de acción frente a impactos ambientales no previstos. También se plantea la necesidad de medir los posibles impactos socioeconómicos en las comunidades pesqueras, turísticas y productivas de Los Choros y Chungungo, y elaborar un plan de desarrollo en armonía con el plan de desarrollo regional. El tribunal destaca la importancia de establecer un mecanismo vinculante para las decisiones del Centro de Estudios y la participación de actores relevantes en su diseño y gobierno corporativo. En cuanto al "Valor Compartido" con las comunidades, el tribunal considera que es importante que los proyectos entreguen beneficios compartidos y reconoce la propuesta de Dominga en este sentido. Sin embargo, señala que esta materia no se encuentra discutida en el proceso de evaluación y que el compromiso de compartir el valor del negocio no puede sobreseer las causales por las cuales el proyecto fue rechazado. En relación a la supuesta insuficiencia de la línea de base del medio marino en las rutas de navegación, el tribunal considera que la información presentada por el proyecto es suficiente y que el Comité de Ministros no hizo un análisis adecuado al respecto. Se destaca que el compendio de conocimiento entregado por el proyecto es robusto y que la campaña de terreno realizada en las rutas de navegación complementa adecuadamente la línea de base del medio marino. En conclusión, el tribunal acoge los cuestionamientos del reclamante y mantiene el rechazo del proyecto Dominga. El Tribunal Ambiental de Chile revisó una alegación presentada por el Titular del proyecto Dominga, en la que se argumentaba que el Comité de Ministros había aplicado criterios diferentes al evaluar los proyectos Dominga y Cruz Grande en relación a la actividad de navegación. El Titular sostenía que ambos proyectos eran similares y debían ser evaluados de manera análoga para respetar el principio de igualdad ante la ley. Sin embargo, el Tribunal determinó que los proyectos no se encontraban en la misma situación jurídica o material, por lo que era razonable aplicar criterios diferentes. Además, se señaló que el proyecto Dominga no había evaluado adecuadamente los impactos de la navegación, a pesar de haber identificado riesgos y presentado planes de contingencia. En conclusión, el Tribunal rechazó la alegación de trato discriminatorio y confirmó la decisión del Comité de Ministros. El Tribunal Ambiental ha analizado los elementos de temporalidad, objetivos y ubicación de los proyectos Cruz Grande y Dominga. Ambos proyectos fueron presentados a través de Estudios de Impacto Ambiental y fueron tramitados en forma paralela durante tres años. Ambos proyectos se ubican en la comuna de La Higuera, Región de Coquimbo. Ambos proyectos tienen como objetivo la recepción, acopio y embarque de sustancias minerales, por lo que es necesario evaluar los impactos ambientales asociados a estas actividades. El Comité de Ministros adoptó un actuar distinto respecto al proyecto Dominga, argumentando la insuficiencia de información sobre las rutas de navegación y los impactos que podrían generar. Sin embargo, el Tribunal considera que el Comité de Ministros contaba con la información necesaria para evaluar adecuadamente el proyecto Dominga. En cuanto a los impactos sinérgicos, el Tribunal concluye que se realizó un análisis completo de estos impactos en el proyecto Dominga, considerando la existencia del proyecto Cruz Grande. El Tribunal considera que la información de línea base es suficiente para definir los impactos sinérgicos y que no existe un desequilibrio entre los impactos y las medidas propuestas. El Tribunal Ambiental ha revisado las reclamaciones presentadas en relación al proyecto minero Dominga. En cuanto a la insuficiencia de la línea de base del medio marino en las rutas de navegación, el Tribunal concluye que no está debidamente acreditada y que las medidas asociadas son adecuadas. En relación al impacto en el recurso hídrico por el cierre del Sistema de Captación y Devolución de aguas desde el Rajo Sur, el Comité de Ministros consideró que la medida de mitigación era inadecuada, pero el SEA argumenta que la medida es idónea y que se puede financiar a través del Fondo para la Gestión de Faenas Mineras Cerradas. Respecto a los efectos en el mar, el Comité de Ministros concluyó que no se evaluaron adecuadamente los impactos de la planta desalinizadora, pero el SEA argumenta que si se evaluaron y que no se evidenciaban cambios significativos en las condiciones físico-químicas ni en la biodiversidad asociada a las Áreas de Manejo y Explotación de Recursos Bentónicos. Además, se establece la necesidad de un monitoreo de la salinidad durante la operación del proyecto. En conclusión, el Tribunal rechaza las reclamaciones presentadas. El Tribunal Ambiental concluye que las preocupaciones presentadas por la empresa Dominga en relación a las preocupaciones PAC (Programa de Adecuación y Compensación) fueron debidamente resueltas. Sin embargo, se determina que hubo una subestimación en la evaluación de los impactos por aporte de material particulado sedimentable (MPS) y se evidencia un desconocimiento de los antecedentes por parte del Comité de Ministros. Además, se señala que la normativa ambiental no se adecúa al contexto del proyecto. En cuanto a la medida de compensación para el impacto en el proceso de nidificación del Pinguino de Humboldt, se determina que la propuesta presentada por la empresa no logra un efecto positivo alternativo y equivalente, por lo que se acoge la reclamación. Se cuestiona la falta de fundamentos del Comité de Ministros y se sugiere la implementación de medidas de mitigación más innovadoras. El Tribunal Ambiental advierte una debilidad en la forma en que se abordó la afectación del Pinguino de Humboldt en el islote Totoralillo Norte durante la evaluación ambiental del proyecto. Se señala que no se consideraron medidas de mitigación innovadoras ni propuestas alternativas de medidas compensatorias. Además, se cuestiona la medida de compensación propuesta, ya que no cumple con los requisitos de equivalencia y adicionalidad. En cuanto al impacto de las barreras eólicas en la avifauna, se argumenta que este tema no fue debidamente considerado ni evaluado durante la evaluación ambiental del proyecto. Se destaca la falta de congruencia y fundamentación en la resolución del Comité de Ministros. Por otro lado, se menciona la importancia de evaluar el efecto atrayente de las luminarias de las barreras eólicas y la infraestructura portuaria en la avifauna marina. Se hace referencia a la contaminación lumínica y sus efectos desorientadores

en las aves costeras.\nLa revisión del expediente de evaluación ambiental no muestra estudios relacionados con la elaboración de una línea base de la avifauna en los sectores mencionados, lo que dificulta la identificación y abordaje de los impactos potenciales. El uso de luminarias durante la noche en zonas portuarias y la mortalidad de aves marinas, especialmente los petreles, son temas sensibles que deben ser considerados. Se advierte la oportunidad de complementar la información técnica para evaluar adecuadamente estos impactos. Además, se cuestiona la supuesta suficiencia de las áreas de exclusión y la falta de consideración de los impactos sinérgicos. Se destaca la falta de estudios sobre el impacto acústico de los buques en la fauna marina. En conclusión, se considera que no se evaluaron adecuadamente los impactos mencionados y se sugiere la realización de investigaciones científicas para incrementar el conocimiento de los ecosistemas marino costeros.\nEl Primer Tribunal Ambiental de Chile ha acogido la reclamación presentada por Andes Iron SpA en contra de la Resolución 1146 del Director Ejecutivo del SEA y el acuerdo N° 08/2017 del Comité de Ministros. El tribunal ha ordenado retrotraer el procedimiento de evaluación ambiental a la etapa posterior al Informe Consolidado de Evaluación Ambiental y realizar una nueva votación ajustada a derecho por parte de la COEVA de la Región de Coquimbo. No se condena en costas a las partes.\nResumen:\nEl 16 de abril de 2021, se notificó en Antofagasta la sentencia anterior a través del estado diario y correo electrónico. La sentencia se encuentra en las fojas 6156.',

```

metadata={
  'caratula': 'Andes Iron SpA con SEA',
  'competencia': 'Art. 17 N° 5 de Ley N° 20.600',
  'componente_ambiental': 'Agua',
  'decision_del_tribunal': 'Acoge',
  'descriptores': 'Titular del proyecto; vicios formales; motivación; nulidad;
vicios de procedimiento; plazos; votación; debido proceso; garantías; contradictoriedad;
congruencias; igualdad ante la ley; medio marino',
  'fechaSentencia': '27-04-2018',
  'nombreProyecto': 'Dominga',
  'page': 0,
  'rol': 'R-1-2017',
  'source':
'/home/sjvasconcello/usm-data/ds-senior/cmad-ai-sustentable-01/etl/load/../../data/clean/docs_
with_summary/R-1-2017.txt',
  'tribunal': 'ITA',
  'urlSentencia':
'https://www.buscadorambiental.cl/buscador-api/file/view/jurisprudencia/SITA Rol N°
R-1-2017?id=1'
},
Document(
  page_content='El caso Caserones\n\nRol: "R-41-2021", Rol: "R-41-2021", Rol:
"R-41-2021", fecha 31-08-2021 de Asociación de Productores y Exportadores Agrícolas del Valle
de Copiapó y otros con Superintendencia de Medio Ambiente cuyo proyecto corresponde a Proyecto
Caserones. Este caso tiene como principal materia Programa de cumplimiento SMA y su componente
ambiental es Agua\n\nThe integration ministers are: \n1. Mauricio Oviedo Gutiérrez\n2. Juan
Opazo Lagos\n3. Marcelo Hernández Rojas\n\nSiendo sus descriptores: Procedimiento
sancionatorio, programa de cumplimiento, SMA, desagregación de cargos, desviación del
procedimiento, integridad del programa de cumplimiento, procedencia del PDC, infracciones que
han causado daño ambiental, criterio de integridad, plan de reparación ambiental, incentivos
al cumplimiento, falta de perjuicio, principio de confianza legítima, calidad de aguas,
acuíferos, criterio de eficacia, razonabilidad y motivación de la resolución de la SMA,
idoneidad de las medidas, Guía para la presentación de Programas de Cumplimiento por
infracciones a instrumentos de carácter ambiental.\n\nEl First Environmental Court Rechaza la
causa en esta ocasión. Los hechos controvertidos son: Procedencia del Programa de Cumplimiento
para infracciones que han causado daño ambiental, Sobre el criterio de eficacia del Programa
de Cumplimiento y la prohibición general de eludir a la responsabilidad.\n\nY el resumen es:
En este documento se presenta una revisión de un caso judicial en el que la Asociación de
Productores y Exportadores Agrícolas del Valle de Copiapó, la Comunidad de Aguas Subterráneas
Sector N° 1 "Aguas Arriba del Embalse Lautaro", la Comunidad de Aguas Subterráneas Sector N° 2
"Embalse Lautaro - La Puerta" y la Comunidad de Aguas Subterráneas Sector N° 3 "La Puerta -
Mal Paso" presentaron una reclamación contra la Resolución Exenta N° 15/Rol D-018-2019 de la
Superintendencia del Medio Ambiente. En la reclamación, solicitan la invalidación de la
resolución y el rechazo del programa de cumplimiento presentado por SCM Minera Lumina Copper
Chile S.A. El caso involucra cargos relacionados con el incumplimiento de condiciones
ambientales establecidas para el proyecto Caserones. La Superintendencia del Medio Ambiente y
SCM Minera Lumina Copper Chile S.A. presentaron sus argumentos en contra de la reclamación. El
Tribunal admitió la reclamación a trámite y rechazó la solicitud de medida cautelar presentada
por los reclamantes. El caso se encuentra en estado de relación y se llevó a cabo una vista el
5 de mayo de 2021.\nLa reclamante argumenta que la resolución impugnada infringe la ley al
desagregar dos de los cargos formulados en la Resolución Exenta N° 1/D-018-2019, lo cual
atenta contra la finalidad de poner término al procedimiento administrativo sancionador a
través del programa de cumplimiento. Además, sostiene que la desagregación de los cargos
infringe el criterio de integridad y que la resolución reclamada carece de una causal legal
que habilite la desagregación del procedimiento administrativo. Por otro lado, la reclamante
acusa a la resolución de violar el principio de confianza legítima al desagregar el
procedimiento sin causa legal alguna. En cuanto a los argumentos de la reclamada, esta
sostiene que la resolución reclamada es legal y fue dictada conforme a la normativa vigente.
Además, argumenta que la empresa ha acreditado la no generación de efectos negativos en
relación al recurso hídrico y que las medidas propuestas en el programa de cumplimiento son
eficaces para contener la pluma de contaminación. También señala que la empresa ha cumplido
con las acciones de difusión y sociabilización del pueblo Colla, pese a no haber sido
coordinadas con la CONADI. En relación al cargo N° 9, la reclamada sostiene que la empresa
pudo acreditar la no generación de efectos negativos y que el caudal entregado de agua
desalada permitió mejorar la calidad del agua entregada a la población. Por último, la
reclamada argumenta que la improcedencia del programa de cumplimiento en casos de daño
ambiental se debe a las características de las infracciones y al contexto regulatorio
ambiental en que se inserten.\nEl Programa de Cumplimiento (PDC) no es un derecho del titular
que deba operar en todo momento, ya que solo cumple su objetivo a través de medidas que
permitan volver al cumplimiento de la normativa ambiental. La desagregación de los cargos 11 y
12 por parte de la Superintendencia del Medio Ambiente (SMA) cumple con la finalidad del
procedimiento sancionatorio de obtener la responsabilidad del infractor y proteger el medio
ambiente. La improcedencia del PDC respecto a estos cargos no genera perjuicios a los
reclamantes, ya que el derecho a presentar un PDC corresponde únicamente al presunto infractor
y no a los interesados del procedimiento. La SMA argumenta que la desagregación de los cargos

```

11 y 12 está en armonía con los intereses de los denunciantes. Además, la SMA se refiere a la Resolución Exenta N° 7/Rol D-011-2015 para justificar su actuar y sostiene que ha sido consistente con la práctica del Servicio. Por otro lado, las reclamantes argumentan que la desagregación de los cargos 11 y 12 vulnera el principio de integridad y que no existe una norma legal que justifique esta desagregación. También sostienen que el PDC y los Planes de Reparación Ambiental (PRA) son mecanismos compatibles y que el cumplimiento satisfactorio de un PDC constituye un equivalente procesal al procedimiento sancionatorio. La SMA argumenta que el PDC no es procedente para infracciones que han causado daño ambiental, ya que no existen acciones que puedan recomponer el medio ambiente dañado a través de este instrumento. Además, señala que el PDC no es un derecho que deba operar a todo evento y que la desagregación de los cargos 11 y 12 no genera perjuicios a los reclamantes. En cuanto al principio de confianza legítima, la SMA sostiene que su actuar ha sido consistente con sus propios precedentes. En resumen, las controversias se centran en la procedencia del PDC para infracciones que han causado daño ambiental y en la compatibilidad entre el PDC y los PRA. La SMA argumenta que el PDC no es procedente para este tipo de infracciones, mientras que las reclamantes sostienen lo contrario. El documento revisa la aplicación de un Programa de Cumplimiento (PDC) y un Plan de Reparación Ambiental (PRA) en el contexto de un procedimiento administrativo sancionatorio por infracciones ambientales. Se destaca que el PDC es un instrumento voluntario de incentivo al cumplimiento normativo, mientras que el PRA se aplica en casos de infracciones con daño ambiental. Se argumenta que la desagregación de los cargos 11 y 12, que generó la aprobación de un PDC y la continuación del procedimiento para el PRA, no vulnera el criterio de integridad y no produce perjuicio a los reclamantes. Además, se señala que la decisión de desagregación se ajusta a la ley y no afecta el principio de confianza legítima. El artículo revisa el principio de confianza legítima en la actuación de la administración como límite a la potestad invalidatoria. Se analiza la actuación de la Superintendencia del Medio Ambiente (SMA) en casos similares para determinar si se ha infringido este principio. Se destaca que la SMA ha establecido un criterio expreso para el caso de infracciones con daño ambiental y su incompatibilidad con el Programa de Cumplimiento (PdC). Se mencionan casos en los que la SMA ha aprobado o rechazado PdC de acuerdo a este criterio. Se discute la eficacia de las medidas incluidas en el PdC y se concluye que la SMA ha realizado un análisis razonable y motivado para descartar los efectos negativos de las infracciones. Se rechaza la reclamación presentada por las asociaciones agrícolas y de aguas subterráneas. Se hace un llamado a la SMA para agilizar la revisión de los cargos desagregados y tomar medidas oportunas en el procedimiento administrativo sancionador en trámite. El Primer Tribunal Ambiental de Chile emitió una sentencia en el caso de la Quebrada La Brea y el Río Ramadillas. En la sentencia, se establece que la empresa LOSMA debe presentar un "Plan de Reparación" por los daños ambientales causados y, en caso de no ser posible, los antecedentes serán derivados al Consejo de Defensa del Estado para que ejerza la acción judicial correspondiente. El tribunal destaca la importancia de garantizar el derecho a vivir en un medio ambiente libre de contaminación y el deber del Estado de preservar los recursos hídricos. La sentencia fue redactada por el Ministro Mauricio Oviedo Gutiérrez y firmada por los Ministros Marcelo Hernández Rojas y Juan Opazo Lagos. El Secretario Abogado del Tribunal, Pablo Miranda Nigro, autorizó la notificación de la sentencia.

```

    'metadatos': {
      'caratula': 'Asociación de Productores y Exportadores Agrícolas del Valle de Copiapó y otros con Superintendencia de Medio Ambiente',
      'competencia': 'Art. 17 N° 3 de Ley N° 20.600',
      'componente_ambiental': 'Agua',
      'decision_del_tribunal': 'Rechaza',
      'descriptores': 'Procedimiento sancionatorio; programa de cumplimiento; SMA; desagregación de cargos; desviación del procedimiento; integridad del programa de cumplimiento; procedencia del PDC; infracciones que han causado daño ambiental; criterio de integridad; plan de reparación ambiental; incentivos al cumplimiento; falta de perjuicio; principio de confianza legítima; calidad de aguas; acuíferos; criterio de eficacia; razonabilidad y motivación de la resolución de la SMA; idoneidad de las medidas; Guía para la presentación de Programas de Cumplimiento por infracciones a instrumentos de carácter ambiental.',
      'fechaSentencia': '31-08-2021',
      'nombreProyecto': 'Proyecto Caserones',
      'page': 0,
      'rol': 'R-41-2021',
      'source': '/home/sjvasconcello/usm-data/ds-senior/cmdd-ai-sustentable-01/etl/load/../../data/clean/docs_with_summary/R-41-2021.txt',
      'tribunal': 'ITA',
      'urlSentencia': 'https://www.buscadorambiental.cl/buscador-api/file/view/jurisprudencia/SITA Rol N° R-41-2021?id=28'
    }
  },
  Document(
    page_content='El caso Escombrera Palguin bajo\n\nRol: "R-25-2019", Rol: "R-25-2019", Rol: "R-25-2019", fecha 24-03-2020 de Elsa Quirquitripay Antiman y Otros con SEREMI de Salud Región de la Araucanía cuyo proyecto corresponde a Escombrera, sitio de disposición de residuos no peligrosos. Este caso tiene como principal materia Invalidación administrativa y su componente ambiental es Agua\n\nThe integration ministers are: \n1. Iván Hunter Ampuero\n2. Jorge Retamal Valenzuela\n3. Sibel Villalobos Volpi\n\nSiendo sus descriptores: Competencia absoluta, instrumento de gestión ambiental, legitimación activa, acción, invalidación facultad, invalidación impropia\n\nEl Third Environmental Court Rechaza la causa en esta ocasión. Los hechos controvertidos son: Si el Tribunal tendría competencia para pronunciarse sobre la reclamación judicial, considerando la naturaleza de las resoluciones impugnadas, Si los Reclamantes tendrían legitimación activa para ejercer legalmente la impugnación judicial\n\nY el resumen es: El abogado Marco Antonio Román Cordero, en representación de ACONSER RESIDUOS SPA, interpuso una reclamación contra la SUPERINTENDENCIA DEL MEDIO AMBIENTE (SMA) por la Resolución Ex. N° 427, que requería a la empresa el ingreso del proyecto "Vertedero Aconser Mocopulli" al Sistema de Evaluación de Impacto Ambiental (SEIA). La reclamante solicitó al tribunal acoger su reclamación y declarar la ilegalidad de la resolución reclamada. Por su parte, la SMA solicitó al tribunal rechazar la reclamación y declarar la legalidad de la resolución. En los antecedentes del acto administrativo reclamado, se mencionan diversos documentos y comunicaciones entre la SEREMI de Salud, la SMA y la Reclamante. En cuanto al procedimiento de reclamación, se detallan las alegaciones de ambas partes y los argumentos presentados. La Reclamante argumenta que el procedimiento de requerimiento de ingreso se inició de manera irregular y que se violaron plazos establecidos por la ley. Además, cuestiona la competencia de la SMA para dictar la

```

resolución reclamada. La SMA, por su parte, argumenta que la documentación relevante está incluida en el expediente administrativo y que la Reclamante tuvo acceso a la información necesaria. También se menciona que el desconocimiento de ciertos documentos no supone indefensión, ya que la SMA constató los mismos hallazgos comunicados por la SEREMI de Salud. En conclusión, la Reclamante solicita la anulación o modificación de la resolución reclamada, mientras que la SMA pide que se declare su legalidad.

La reclamación presentada por ACONSER Residuos SpA no cuestiona el cumplimiento de los supuestos fácticos y jurídicos contenidos en la resolución impugnada. Sin embargo, se plantean controversias relacionadas con el procedimiento administrativo que precedió a la dictación del acto terminal de requerimiento de ingreso al SEIA. Estas controversias incluyen la falta de conocimiento de la denuncia que originó el requerimiento, la demora excesiva en resolver el procedimiento, la omisión en la aplicación del Instructivo para la Tramitación de Requerimientos de Ingreso al SEIA y la omisión en decretar diligencias probatorias solicitadas por la interesada. Respecto a la falta de conocimiento de la denuncia, se determina que el expediente administrativo físico contiene el procedimiento que justifica la dictación del acto, por lo que la tardanza en incorporar la denuncia al sistema electrónico no constituye un vicio de legalidad. En cuanto a la demora en resolver el procedimiento, se establece que el plazo de 6 meses establecido por la ley no se ha superado. Además, se señala que el incumplimiento de este plazo no afecta la validez del procedimiento, sino que compromete la responsabilidad funcionaria. En relación a la omisión en la aplicación del Instructivo, se argumenta que este no establece la obligatoriedad de seguir el orden de tramitación indicado. Por último, se concluye que las solicitudes de diligencias probatorias solicitadas por la interesada no son necesarias para confirmar o rechazar la verificación de los hechos que configuran la elusión atribuida en la resolución impugnada. En resumen, se determina que las controversias planteadas por la reclamante no afectan la legalidad del procedimiento administrativo.

En resumen, el caso trata sobre una reclamación presentada por un titular de un proyecto contra la Superintendencia del Medio Ambiente (SMA) en relación con el requerimiento de ingreso al Sistema de Evaluación de Impacto Ambiental (SEIA). El reclamante alega que la SMA infringió el Instructivo para la Tramitación de Requerimientos de Ingreso al SEIA al no solicitar previamente el informe del Servicio de Evaluación Ambiental (SEA) antes de iniciar el procedimiento. Sin embargo, la SMA argumenta que cumplió con lo establecido en la ley y que el Instructivo no es de carácter obligatorio. El tribunal concluye que la SMA actuó de acuerdo a la ley y que no se produjo indefensión para el reclamante. Además, se analiza la omisión de decretar las diligencias probatorias solicitadas por el interesado y se determina que no causaron indefensión y que eran innecesarias para el caso. En consecuencia, el tribunal considera que la resolución de ingreso al SEIA se ajusta a derecho.

El Tribunal Ambiental rechazó la reclamación presentada por la impugnante en contra de la Resolución Exenta N° 427 de la Superintendencia del Medio Ambiente (SMA). La impugnante solicitaba la realización de diligencias probatorias, pero no especificó de manera clara la conexión entre dichas diligencias y los hechos que deseaba probar. Además, no indicó las consecuencias jurídicas que se derivarían de la falta de pronunciamiento de la autoridad sectorial respecto a los recursos administrativos pendientes de resolución. Por lo tanto, el Tribunal consideró que la Resolución de la SMA se ajusta a derecho. No se condenó en costas a la impugnante por considerar que litigó con motivos plausibles.

```

    metadata={
      'caratula': 'Asociación Indígena Aymara Salar de Coposa con SMA ',
      'competencia': 'Art. 17 N° 3 de Ley N° 20.600',
      'componente_ambiental': 'Agua',
      'decision_del_tribunal': 'Acoge',
      'descriptores': 'Procedimiento sancionatorio; programa de cumplimiento;
    consulta indígena; recurso hídrico; daño ambiental; acuíferos; criterios; integridad;
    eficacia; verificabilidad; derecho a los recursos naturales; principio de participación;
    motivación; reincidencia ',
      'fechaSentencia': '30-12-2019',
      'nombreProyecto': 'Collahuasi',
      'page': 0,
      'rol': 'R-25-2019',
      'source':
        '/home/sjvasconcello/usm-data/ds-senior/cmdad-ai-sustentable-01/etl/load/../../data/clean/docs_
    with_summary/R-25-2019.txt',
      'tribunal': 'ITA',
      'urlSentencia':
        'https://www.buscadorambiental.cl/buscador-api/file/view/jurisprudencia/SITA Rol N°
    R-25-2019?id=16'
    }
  ),
  Document(
    page_content='El caso Pascua Lama\n\nRol: "R-5-2018", Rol: "R-5-2018",Rol:
    "R-5-2018", fecha 17-09-2020 de Compañía Minera Nevada SpA con Superintendencia del Medio
    Ambiente cuyo proyecto corresponde a Pascua Lama. Este caso tiene como principal materia
    Sancionatorio ambiental y su componente ambiental es Agua\n\nThe integration ministers are:
    \n1. Cristián Delpiano Lira\n2. Jasna Pavlich Núñez\n3. Fabrizio Queirolo Pellerano\n\nSiendo
    sus descriptores: Sancionatorio ambiental, estándar de motivación - proporcionalidad -
    principio de non bis in idem - decaimiento - metodología de determinación de sanciones -
    legitimación activa - tasa de descuento - daño ambiental - impacto ambiental - clasificación
    de la infracción - circunstancias del art. 40 de la Ley N° 20.417\n\nEl First Environmental
    Court Acoge parcialmente la causa en esta ocaion. Los hechos controvertidos son: Ilegalidad
    transversal de las sanciones impuestas por la Res. Ex. N°72/2018. De las supuestas
    ilegalidades respecto de las sanciones de clausura definitiva aplicadas a cinco de los cargos
    formulados. De las supuestas ilegalidades cometidas en la Res. Ex. N°72/2018, respecto de las
    sanciones de multa aplicadas, al resto de los cargos formulados. Del supuesto exceso en las
    competencias que habría incurrido la SMA respecto de la manera en que pretende que se
    implemente la clausura. Respecto a la falta de legitimación activa alegada por la SMA en
    contra de Agrícola Dos Hermanos y Agrícola Santa Mónica Limitada. Respecto de la Resolución
    Exenta N°70. Respecto de la Res. Ex. N°72. Legitimación activa de los reclamantes de la
    Reclamación signada con el rol R-6-2018. Estándar de motivación exigido. Resolución Exenta
    N°70, de 2018, de la Superintendencia del Medio Ambiente. Decaimiento del procedimiento
    administrativo y la falta de oportunidad en el ejercicio de la potestad sancionatorio. De la
    supuesta infracción al principio non bis in idem. En cuanto a la metodología para la
    determinación de sanciones ambientales, la tasa de descuento, y la proporcionalidad de las
    sanciones, de la Resolución Exenta N°72/2018.2.7 En cuanto al cálculo de la Tasa de Descuento
    descrita en las "Bases Metodológicas para Determinación de Sanciones Ambientales". El estándar
    de proporcionalidad en el ejercicio de la potestad sancionatoria de la Superintendencia del
    Medio Ambiente. III. DEL ANÁLISIS DE LAS RECLAMACIONES EFECTUADAS PARA CADA CARGO EN
    PARTICULAR ALEGADO RESPECTO DE LA RESOLUCIÓN EXENTA N°72, REALIZADO EN CAUSA R-5-2018. 3.1 De
  
```


las supuestas ilegalidades cometidas en la sanción de clausura definitiva aplicada al cargo 23.2. 3.2 De las supuestas ilegalidades cometidas en la sanción de clausura definitiva aplicada al cargo 23.9. 3.3 De las supuestas ilegalidades cometidas en la sanción de clausura definitiva aplicada al cargo 23.11. 3.4 De las supuestas ilegalidades cometidas en la sanción de clausura definitiva aplicada al cargo 4. 3.5 De las supuestas ilegalidades cometidas en la sanción de clausura definitiva aplicada al cargo N°7. 3.6 De las supuestas ilegalidades cometidas en la sanción de multa aplicada al cargo 23.8. 3.7 De las supuestas ilegalidades cometidas en la sanción de multa aplicada al cargo 6. 3.8 De las supuestas ilegalidades cometidas en la definición de todas las sanciones de multa al establecer un criterio de tasa de descuento que es muy superior al de las empresas similares. 3.9 Del supuesto exceso en las competencias que habría incurrido la SMA respecto de la manera en que pretende que se implemente la clausura. IV. DEL ANÁLISIS DE LAS RECLAMACIONES EFECTUADAS PARA CADA CARGO EN PARTICULAR ALEGADO RESPECTO DE LA RESOLUCIÓN EXENTA N°72, REALIZADO EN CAUSA R-6-2018. 4.1 Vicio N° 1: Respecto a la omisión de la resolución sancionatoria de ordenar a CMN SpA la construcción definitiva del Sistema de Manejo de Aguas. 4.2 Vicio N°2: Sobre la falta de contenido sancionatorio y disuasivo en la aplicación de 5 sanciones de clausura. 4.3 Vicio N°3: Respecto al descarte ilegal de daño ambiental en las aguas. 4.4 Vicio N°4: De la supuesta falta de fundamentación de la resolución reclamada en lo que dice relación con la causalidad en la contaminación de las aguas. 4.5 Vicio N°5. Respecto de la motivación de la SMA para recalificar y absolver a CMN SpA de los cargos 23.13, 24.2 y 25. 4.6 Vicio N°6: Sobre la eventual falta de proporcionalidad en la aplicación de las sanciones. 4.7 Vicio N°7: La resolución recurrida no pondera ni aplica en concreto las circunstancias que contempla el artículo 40 de la LOSMA para efectos de determinar la sanción a aplicar a cada infracción. 4.8 Del supuesto problema del petitorio de la reclamación.',

```

metadata={
  'caratula': 'Compañía Minera Nevada SpA con Superintendencia del Medio
Ambiente',
  'competencia': 'Art. 17 N° 3 de Ley N° 20.600',
  'componente_ambiental': 'Agua',
  'decision_del_tribunal': 'Acoge parcialmente',
  'descriptores': 'Sancionatorio ambiental; estándar de motivación -
proporcionalidad - principio de non bis in idem - decaimiento - metodología de determinación
de sanciones - legitimación activa - tasa de descuento - daño ambiental - impacto ambiental -
clasificación de la infracción - circunstancias del art. 40 de la Ley N° 20.417',
  'fechaSentencia': '17-09-2020',
  'nombreProyecto': 'Pascua Lama',
  'page': 0,
  'rol': 'R-5-2018',
  'source':
'/home/sjvasconcello/usm-data/ds-senior/cmdad-ai-sustentable-01/etl/load/../../data/clean/docs_
with_summary/R-5-2018.txt',
  'tribunal': 'ITA',
  'urlSentencia':
'https://www.buscadorambiental.cl/buscador-api/file/view/jurisprudencia/SITA Rol N°
R-5-2018?id=20'
}
)
]
}
\end{table}

```