UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Ingeniería

GPT *FINE-TUNING* PARA SU ESPECIALIZACIÓN EN LAS LEYES DE PROPIEDAD INTELECTUAL EN GUATEMALA

Trabajo de graduación presentado por Oliver Josué de León Milian para optar al grado académico de Licenciado en Ingeniería en Ciencias de la Computación y Tecnologías de la Información

Guatemala,

2024

UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Ingeniería

GPT *FINE-TUNING* PARA SU ESPECIALIZACIÓN EN LAS LEYES DE PROPIEDAD INTELECTUAL EN GUATEMALA

Trabajo de graduación presentado por Oliver Josué de León Milian para optar al grado académico de Licenciado en Ingeniería en Ciencias de la Computación y Tecnologías de la Información

Guatemala,

2024

Vo.Bo.:

(f) _____

MSc. Luis Alberto Suriano Saravia

Tribunal examinador:

(f) _____ Giancarlo Eutoshenko Figueroa Figueroa

2410 28981 0101

Lic. Giancarlo E. Figueroa Figueroa

(f) _____

MSc. Luis Alberto Suriano Saravia

(f) _____

MSc. Douglas Leonel Barrios Gonzalez

Fecha de aprobación del examen de graduación:

Guatemala, 17 de Junio de 2024

Índice

Índice de cuadros

Índice de figuras

Resumen

En el entorno complejo del derecho guatemalteco, encontrar respuestas claras y rápidas puede ser sumamente desafiante. La abundancia de leyes y regulaciones crea un laberinto informativo para los profesionales legales. Este proyecto surge como respuesta a este desafío, utilizando Modelos de Lenguaje de Aprendizaje Profundo (Large Language Model - LLM por sus siglas en inglés) para proporcionar orientación en este laberinto legal. La justificación se encuentra en los obstáculos que enfrentan los profesionales del derecho, como la falta de acceso a fuentes digitales y la complejidad del sistema legal. Mediante el *fine-tuning* de modelos de lenguaje natural, se adapta un modelo para comprender y responder consultas sobre la ley de propiedad intelectual. Los resultados muestran mejoras significativas en la capacidad del modelo para abordar consultas legales. Además, se implementa un sistema de gestión de consultas para facilitar la interacción usuario-tecnología. Este proyecto aspira no solo a mejorar el acceso a la información legal, sino también a fortalecer la práctica legal en Guatemala, proporcionando orientación y claridad en este laberinto legal. Este proceso representa un camino hacia la creación de un sistema legal más accesible, eficiente y equitativo.

Abstract

In the complex environment of Guatemalan law, finding clear and quick answers can be extremely challenging. The abundance of laws and regulations creates an informational maze for legal professionals. This project arises as a response to this challenge, utilizing Deep Learning Language Models (LLMs) to provide guidance in this legal labyrinth. The justification lies in the obstacles faced by legal professionals, such as the lack of access to digital sources and the complexity of the legal system. Through fine-tuning of natural language models, a model is adapted to understand and respond to queries about intellectual property law. The results show significant improvements in the model's ability to address legal queries. Additionally, a query management system is implemented to facilitate user-technology interaction. This project aspires not only to improve access to legal information but also to strengthen legal practice in Guatemala, providing guidance and clarity in this legal maze. This is a journey towards a more accessible, efficient, and fair legal system.

1. Introducción

El presente trabajo se centra en la implementación y evaluación de un sistema de gestión de consultas legales basado en el *fine-tuning* de modelos de lenguaje natural, con un enfoque específico en el ámbito de la ley de propiedad intelectual en Guatemala.

El objetivo principal de esta investigación es explorar el potencial del *fine-tuning* de modelos de lenguaje, utilizando el modelo GPT-3.5-turbo-0125, para mejorar la capacidad de comprensión y respuesta a consultas legales especializadas. A través de este enfoque, se pretende ofrecer una solución tecnológica efectiva para la atención automatizada de consultas legales, contribuyendo así a la eficiencia y accesibilidad en el acceso a la información jurídica.

Para alcanzar este objetivo, se emplea una metodología que abarca varias etapas, comenzando con la recolección y etiquetado de datos relevantes en el contexto de la ley de propiedad intelectual en Guatemala. Posteriormente, se procede con la experimentación del *fine-tuning* del modelo de lenguaje, explorando diferentes configuraciones de épocas de entrenamiento para determinar la óptima para la tarea específica.

La implementación del sistema de gestión de consultas implica el desarrollo de una interfaz que permita a los usuarios formular preguntas y recibir respuestas proporcionadas por el modelo ajustado (*fine-tuned*). Además, se integran algoritmos diseñados para asistir al usuario en la formulación adecuada de consultas y en la interpretación de las respuestas obtenidas.

Las conclusiones obtenidas a partir de este estudio proporcionan una visión clara sobre la efectividad del *fine-tuning* de modelos de lenguaje para mejorar la comprensión y respuesta automatizada a consultas legales especializadas. Asimismo, se identifican áreas de mejora y se ofrecen recomendaciones para futuras investigaciones en este campo, con el objetivo de seguir avanzando en la aplicación de tecnologías de procesamiento del lenguaje natural en el ámbito legal

2. Justificación

La complejidad inherente del campo legal y la dificultad para acceder a distintas fuentes jurídicas representan un problema significativo en Guatemala. El área legal se caracteriza por su amplio espectro de leyes, reglamentos, convenios y precedentes legales, lo que dificulta la realización eficiente de procesos y la consulta ágil de información. Esto crea la necesidad de desarrollar herramientas tecnológicas que orienten la obtención de respuestas legales rápidas y exactas para los profesionales del derecho.

En muchos casos, la consulta a la información legal puede ser limitada o ineficiente debido a la inexistencia de copias digitales, la dispersión de fuentes y la complejidad de los sistemas legales. Esto plantea desafíos para aquellos que buscan respuestas legales, lo que subraya la necesidad de desarrollar una solución que supere estas barreras y mejore el acceso a la información legal de manera efectiva.

El avance de los LLMs ofrece una oportunidad prometedora para abordar estas dificultades. Estos modelos de lenguaje han demostrado habilidades significativas en comprensión y generación de contenido en lenguaje natural. Al entrenar y mejorar los LLMs con datos legales específicos del país objetivo, es posible desarrollar una tecnología que genere respuestas capaces de orientar con exactitud y contexto las consultas legales planteadas por profesionales del derecho.

La resolución de consultas legales de manera eficiente y ágil es una necesidad creciente en el campo legal. Los profesionales del derecho buscan soluciones que les permitan obtener respuestas claras y fundamentadas de manera rápida y confiable. La tecnología especializada basada en LLMs propuesta en este proyecto tiene como objetivo agilizar el proceso de obtención de respuestas legales exactas y actualizadas, reduciendo la dependencia de la búsqueda manual y la revisión exhaustiva de fuentes legales.

Mejorar la experiencia y calidad en la asesoría legal es otro factor crucial que motiva este proyecto. Al proporcionar una herramienta basada en LLMs, se busca fortalecer el ejercicio de la profesión legal al permitir a los profesionales del derecho acceder a respuestas claras y fundamentadas, contribuyendo así a una mejor experiencia para los usuarios que requieren asesoramiento legal. Por lo tanto, este proyecto se considera relevante y necesario para abordar los desafíos existentes en el campo legal guatemalteco.

3. Objetivos

3.1. Objetivo general

Orientar al usuario en consultas sobre la ley de propiedad intelectual a través de una solución tecnológica basada en LLM *Fine-Tuning.*

3.2. Objetivos específicos

- Realizar un LLM *Fine-Tuning* usando el modelo GPT para darle la capacidad de entender y responder consultas legales en torno a la ley de propiedad intelectual.

- Implementar un sistema que permita y guíe al usuario en la creación de prompts y la interpretación de las respuestas obtenidas.

4. Marco teórico

4.1. *Large Language Model Fine-Tuning (LLM fine-tuning)*

4.1.1. *Large Language Model (LLM)*

Un modelo de Lenguaje de gran tamaño (*Large Language Model* - LLM por sus siglas en inglés) es un modelo computacional destacado por su capacidad para lograr la generación de lenguaje de propósito general y otras tareas de procesamiento de lenguaje natural, como la clasificación, reconocimiento, sintetización, traducción, predicción y generación de contenido utilizando conjuntos de datos muy grandes.

Los LLM adquieren estas habilidades al aprender relaciones estadísticas en datos secuenciales a partir de documentos de texto durante un proceso de entrenamiento auto-supervisado y semi-supervisado computacionalmente intensivo. Estos representan en gran medida una clase de arquitectura de aprendizaje profundo llamada redes transformadoras. (NVIDIA, 2023)

Las redes transformadoras constituyen una arquitectura específica de redes neuronales diseñada para transformar una secuencia de entrada en una secuencia de salida. Esta transformación se realiza mediante el aprendizaje del contexto y la captura de las relaciones entre los elementos de la secuencia. El modelo de transformador emplea una representación matemática interna que identifica la relevancia y las interacciones entre las palabras para producir el resultado deseado (Amazon, 2023).

LLMs conocidos en la industria:

- ***Large Language Model Meta AI (LLAMA)***

  LLaMA es una familia de modelos de lenguaje autoregresivos (LLMs) desarrollada por Meta AI a partir de febrero de 2023. Utiliza la arquitectura *transformer*, que se ha convertido en el estándar para el modelado del lenguaje desde 2018. Aunque comparte la misma arquitectura básica que GPT-3, LLaMA presenta algunas diferencias arquitectónicas menores, como el uso de funciones de activación no lineales distintas, las representaciones vectoriales de palabras, frases o documentos en un espacio numérico de dimensiones reducidas utilizan posiciones rotativos (estáticas y absolutas) en lugar de posicionales absolutos y cambio en la función de normalización. Además, LLaMA amplía la longitud del contexto de 2,000 (en Llama 1) a 4,000 tokens (en Llama 2) entre los modelos (Meta, 2023).

  LLaMA incluye modelos con diferentes números de parámetros, que van desde 7 mil millones hasta 65 mil millones en la primera versión. En una versión posterior, se lanzaron modelos adicionales con 7 mil millones, 13 mil millones y 70 mil millones de parámetros.

  LLaMA tiene una amplia gama de aplicaciones en inteligencia artificial, que van desde el procesamiento de lenguaje natural hasta la generación de texto y la traducción de idiomas. Además, se puede utilizar para crear chatbots, asistentes virtuales y otros sistemas de inteligencia artificial que dependen de la comunicación en lenguaje natural. Su capacidad para realizar múltiples tareas lingüísticas con alta precisión lo convierte en una solución ideal para empresas que necesitan procesar grandes cantidades de datos textuales y extraer información relevante.

- ***Bidirectional Encoder Representations from Transformers (Bert)***

  Bidirectional Encoder Representations from Transformers (BERT) es un modelo de lenguaje basado en la arquitectura *transformer*, notable por su dramática mejora sobre los modelos previos de vanguardia. Fue introducido en octubre de 2018 por investigadores de Google.

  La arquitectura de BERT es una red codificadora *transformer* bidireccional de múltiples capas, que guarda similitudes con el modelo *transformer*. La estructura *transformer* consiste en una red codificador-decodificador que emplea autoatención en el lado del codificador y atención en el lado del decodificador. También cuenta con redes de retroalimentación más grandes (768 y 1024 unidades ocultas respectivamente) y más cabezas de atención (12 y 16 respectivamente) que la arquitectura *Transformer* original. BERT contiene 110 millones de parámetros (GeeksforGeeks, 2023).

  BERT tiene una amplia variedad de aplicaciones en el procesamiento del lenguaje natural. Se utiliza para generar representaciones de texto, como representaciones vectoriales de palabras, frases o documentos en un espacio numérico de dimensiones reducidas de palabras, lo que facilita la comprensión del contexto y la semántica del texto. Además, BERT se adapta para tareas específicas como el reconocimiento de entidades nombradas y la clasificación de texto, como análisis de sentimientos y detección de . También se emplea en sistemas de preguntas y respuestas, traducción automática, síntesis de texto y sistemas de conversación. Sus representaciones vectoriales se utilizan para medir la similitud semántica entre oraciones o documentos.

- ***Generative Pre-trained Transformer 3.5 (GPT-3.5)***

  El *Generative Pre-trained Transformer 3.5 (GPT-3.5)*, desarrollado por OpenAI en 2022, es una derivación de los modelos GPT-3. Su arquitectura, fundamentada en redes neuronales, se caracteriza por manejar una considerable cantidad de datos textuales, alcanzando los 570 GB, lo que lo posiciona como uno de los modelos de lenguaje más amplios conocidos hasta la fecha (Koshti H., 2023). Con 175 mil millones de parámetros, el GPT-3.5 muestra una capacidad excepcional para abordar diversas tareas lingüísticas, como traducción de idiomas, completado de texto y resolución de preguntas.

  La estructura del GPT-3.5 se basa en la red neuronal transformadora, un tipo de modelo de aprendizaje profundo que ha revolucionado el procesamiento de lenguaje natural. Esta arquitectura consiste en una serie de capas codificadoras y decodificadoras que se entrenan con grandes volúmenes de datos para identificar patrones lingüísticos subyacentes. El GPT-3.5 lleva este enfoque un paso más allá al implementar técnicas avanzadas como el pre-entrenamiento no supervisado y el *fine-tuning*, lo que resulta en un modelo de lenguaje altamente preciso y adaptable. Este entrenamiento diversificado en una variedad de tareas lingüísticas proporciona al GPT-3.5 una comprensión profunda de las complejidades del lenguaje humano.(OpenAI, 2023b)

  La arquitectura GPT-3.5 tiene una amplia gama de aplicaciones en inteligencia artificial, desde procesamiento de lenguaje natural hasta generación de texto y traducción de idiomas. También se puede utilizar para crear *chatbots*, asistentes virtuales y otros sistemas de inteligencia artificial que dependen de la comunicación en lenguaje natural. Su capacidad para realizar múltiples tareas lingüísticas con alta precisión lo convierte en una solución ideal para empresas que necesitan procesar grandes cantidades de datos textuales y extraer información relevante.

4.1.2. Definición: *Large Language Model Fine-Tuning (LLM fine-tuning)*

El *Fine Tuning* de un *Large Language Model* (LLM) implica ajustar sus parámetros para adaptarlo a una tarea específica mediante el entrenamiento con un conjunto de datos relevante. La cantidad

de ajuste necesaria varía según la complejidad de la tarea y el tamaño del conjunto de datos utilizado, lo cual puede implicar múltiples iteraciones a través de épocas para refinar el modelo y mejorar su desempeño en la tarea objetivo (Khawaja, R., 2023).

En el contexto de los modelos de lenguaje, los parámetros representan los valores numéricos que definen las conexiones entre las neuronas en sus distintas capas. Durante el entrenamiento, estos parámetros se ajustan continuamente para minimizar una función de pérdida, que mide la discrepancia entre las predicciones del modelo y los datos reales (Lark Editorial Team, 2023a). El número de parámetros en un modelo es un factor crucial que determina su capacidad para comprender y representar la complejidad del lenguaje natural.

Durante el *Fine Tuning* de *Large Language Models* (LLM), las épocas (*epochs*) son unidades esenciales de progreso en el proceso de ajuste del modelo para tareas específicas. Cada época constituye una iteración completa a través del conjunto de datos de entrenamiento, permitiendo al modelo ajustar gradualmente sus parámetros para minimizar la función de pérdida y mejorar su desempeño (Simplilearn, 2023). La determinación del número óptimo de épocas se basa en consideraciones como la complejidad de la tarea, la diversidad del conjunto de datos y la arquitectura del modelo, con el fin de alcanzar un equilibrio entre la capacidad de generalización y el riesgo de sobreajuste.

Técnicas populares de *fine-tuning* para LLMs:

- ***Fine-tuning* sin supervisión (*unsupervised fine-tuning*)**

  El *fine-tuning* no supervisado es una técnica en la que se entrena el LLM en un conjunto de datos que no contiene etiquetas. Esto significa que el modelo no sabe cuál es la salida correcta para cada entrada. En cambio, el modelo aprende a predecir el siguiente token en una secuencia o a generar texto similar al texto en el conjunto de datos.El *fine-tuning* no supervisado es una técnica de (*fine-tuning*) menos costosa computacionalmente que el (*fine-tuning*) supervisado. Sin embargo, también es menos probable que alcance el mismo nivel de rendimiento (Khawaja, R., 2023).

  Existen diferentes opciones para el (*fine-tuning*) no supervisado. Uno de los métodos de preentrenamiento más eficaces es el *Transformer* y el *Autoencoder* Secuencial de Desruido (*Transformer(-based) and Sequential Denoising Auto-Encoder* - TSDAE por sus siglas en inglés), desarrollado por Kexin Wang, Nils Reimers e Iryna Gurevych en 2021 (Pinecone, 2022).
  TSDAE introduce ruido en las secuencias de entrada al eliminar o intercambiar tokens (por ejemplo, palabras). Estas oraciones dañadas son codificadas por el modelo *transformer* en vectores de oraciones. Luego, otra red decodificadora intenta reconstruir la entrada original a partir de la codificación de la oración dañada.
  A primera vista, esto puede parecer similar al modelado de lenguaje enmascarado (*masked-language modeling* - MLM por sus siglas en inglés). MLM es el enfoque de preentrenamiento más común para los modelos *transformer*. Un número aleatorio de tokens se enmascara utilizando un 'token de enmascaramiento', y el *transformer* debe intentar adivinar qué falta.

- **Aprendizaje por refuerzo usando retroalimentación humana (RLHF por sus siglas en inglés)**

  RLHF, o retroalimentación de aprendizaje por reforzamiento humano, es una técnica en la cual se emplea la retroalimentación humana para ajustar finamente el LLM. La idea básica consiste en proporcionar al LLM un estímulo y este genera una salida. Posteriormente, se solicita a un humano que califique la salida. Esta calificación se utiliza como señal para ajustar finamente el LLM y generar salidas de mayor calidad (Khawaja, R., 2023).

  El algoritmo utilizado para el *fine-tuning* mediante aprendizaje por refuerzo en los modelos

de lenguaje es la Optimización de Política Proximal (*Proximal Policy Optimization* - PPO, por sus siglas en inglés). Este método implica ajustar algunos o todos los parámetros de un modelo de lenguaje con un algoritmo de gradiente de política basado en proximidad. Algunos parámetros del modelo se mantienen fijos debido a que el *fine-tuning* de un modelo completo con un gran número de parámetros sería prohibitivamente costoso. La dinámica exacta de cuántos parámetros congelar o no se considera un problema de investigación abierto. El PPO maximiza las métricas de recompensa en el lote de datos actual, utilizando restricciones en el gradiente para asegurar que el paso de actualización no desestabilice el proceso de aprendizaje (Lambert, N., 2022).

- ***Fine-tuning* supervisado (*supervised fine-tuning*)**

  En este método, el modelo se entrena en un conjunto de datos etiquetados específico de la tarea, donde cada entrada está asociado con una respuesta o etiqueta correcta. El modelo aprende a ajustar sus parámetros para predecir estas etiquetas con la mayor precisión posible. Este proceso orienta al modelo a aplicar su conocimiento preexistente, adquirido del pre-entrenamiento en un conjunto de datos extenso, a la tarea específica en cuestión (SuperAnnotate, 2024). El *fine-tuning* supervisado puede mejorar significativamente el rendimiento del modelo en la tarea, convirtiéndolo en un método eficaz y eficiente para personalizar LLMs.
  El *fine-tuning* supervisado opera sobre la premisa de aprovechar las características previamente aprendidas de un modelo de lenguaje pre-entrenado y adaptarlas para satisfacer tareas específicas. Antes de la ejecución del *fine-tuning*, el modelo pre-entrenado contiene representaciones lingüísticas generales aprendidas a partir de grandes cantidades de datos sin etiquetar. Sin embargo, estas representaciones pueden no ser óptimas para una tarea específica. Durante el proceso de *fine-tuning*, se identifican capas o parámetros particulares del modelo que se ajustarán para optimizar su desempeño en la nueva tarea (Lark Editorial Team, 2023b).

  A través de la optimización basada en gradientes, estos parámetros se actualizan iterativamente durante el entrenamiento para minimizar la pérdida asociada con la tarea específica. Después de la ejecución del *fine-tuning*, el modelo adaptado contiene representaciones más refinadas y especializadas que están mejor alineadas con los requisitos de la tarea objetivo. Además, para evitar el sobreajuste y mejorar la generalización del modelo, se pueden emplear técnicas de regularización, como la regularización L1/L2 o la deserción. Esta adaptación fina no solo permite al modelo integrar conocimientos especializados relevantes para la tarea en cuestión, sino que también mantiene y refina las representaciones de alto nivel aprendidas durante el pre-entrenamiento.

  Buenas prácticas para desarrollar *fine-tuning*:

- **Preparación de datos**

  La preparación de datos implica la curación y preprocesamiento del conjunto de datos para garantizar su relevancia y calidad para la tarea específica. Además de estas tareas fundamentales, se emplean técnicas de limpieza de datos para abordar aspectos como la eliminación de datos duplicados, el manejo de valores faltantes, la corrección de errores de formato y la identificación y tratamiento de valores atípicos (Turing, 2023). Estas prácticas aseguran que los datos estén libres de inconsistencias y ruidos, lo que contribuye a un entrenamiento más efectivo y preciso del modelo.

- **Selección del modelo**

  Es crucial seleccionar un modelo pre-entrenado que se alinee con los requisitos específicos de la tarea o dominio objetivo. Comprender la arquitectura, las especificaciones de entrada/salida y las capas del modelo pre-entrenado es esencial para una integración fluida en el flujo de trabajo

de *fine-tuning*.

Factores como el tamaño del modelo, los datos de entrenamiento y el rendimiento en tareas relevantes deben considerarse al tomar esta decisión (Multimodal, 2023). Al seleccionar un modelo pre-entrenado que se ajuste estrechamente a las características de la tarea objetivo, se puede agilizar el proceso de *fine-tuning* y maximizar la adaptabilidad y efectividad del modelo para la aplicación prevista.

- **Validación**

  La validación implica evaluar el rendimiento de un *fine-tuning* utilizando un conjunto de validación. Monitorear métricas como exactitud, pérdida, precisión y recuperación proporciona información sobre la efectividad del modelo y sus capacidades de generalización.

  - **Exactitud**: Esta métrica indica la proporción de predicciones correctas realizadas por el modelo con respecto al total de predicciones realizadas (Google, 2022a). En el *fine-tuning* de un LLM, la precisión es esencial para medir qué tan bien el modelo clasifica o genera resultados de acuerdo con las etiquetas o respuestas esperadas en el conjunto de datos de entrenamiento o validación.

  - **Pérdida**: La pérdida, también conocida como función de pérdida, representa la discrepancia entre las predicciones del modelo y las respuestas reales en el conjunto de datos de entrenamiento (Google, 2022b). Durante el fine-tuning, el objetivo es minimizar esta pérdida para que el modelo se ajuste mejor a los datos y haga predicciones más precisas.

  - **Precisión y recuperación**: La precisión y la recuperación son métricas fundamentales en la evaluación de sistemas, especialmente en contextos donde se busca recuperar información relevante, como en la generación de respuestas a preguntas. La precisión se define como la proporción de respuestas generadas que son correctas y pertinentes con respecto al total de respuestas generadas. Por otro lado, la recuperación se refiere a la proporción de respuestas relevantes que el sistema ha identificado correctamente en relación con todas las respuestas relevantes disponibles en el conjunto de datos (Saxena S., 2018). Ambas métricas son de suma importancia para garantizar que el sistema no solo proporcione respuestas precisas, sino que también sea capaz de identificar y recuperar la mayor cantidad posible de información relevante en el conjunto de prueba.

  Al evaluar estas métricas, se puede medir qué tan bien el modelo ajustado finamente está desempeñándose en los datos específicos de la tarea e identificar áreas potenciales de mejora. Este proceso de validación permite el refinamiento de los parámetros de *fine-tuning* y la arquitectura del modelo, lo que finalmente conduce a un modelo optimizado que sobresale en la generación de resultados precisos para la aplicación prevista.

4.2. Introducción a la propiedad intelectual en Guatemala

4.2.1. Propiedad intelectual

La propiedad intelectual se refiere a los derechos legales y exclusivos que se otorgan sobre creaciones de la mente humana. Estas creaciones pueden ser tanto intangibles como tangibles, y abarcan una amplia gama de áreas, incluyendo obras literarias, artísticas, invenciones, símbolos, nombres, imágenes y diseños utilizados en el comercio (World Intellectual Property Organization, 2020). La propiedad intelectual se divide generalmente en dos categorías principales: derechos de autor y derechos de propiedad industrial.

Derechos de autor

Los derechos de autor protegen obras de autoría, como libros, música, películas, obras de arte y software, otorgando al creador el derecho exclusivo de reproducir, distribuir y mostrar públicamente su obra.

Derechos de propiedad industrial

Los derechos de propiedad industrial son un conjunto de derechos legales que protegen las creaciones intangibles relacionadas con la actividad industrial y comercial (**15**). Estos derechos otorgan a sus titulares exclusividad sobre sus invenciones o creaciones, permitiéndoles controlar su uso y explotación por parte de terceros. Los derechos de propiedad industrial se dividen principalmente en cuatro categorías:

- **Patentes:** Protegen las invenciones técnicas, como productos, procesos o mejoras técnicas, otorgando a su titular el derecho exclusivo a explotar la invención durante un período determinado de tiempo.
- **Marcas:** Protegen los signos distintivos utilizados para identificar productos o servicios en el mercado, como nombres, logotipos, símbolos o *slogans*, garantizando que solo el titular tenga derecho a usarlos en relación con los productos o servicios designados.
- **Diseños industriales:** Protegen la apariencia estética de un producto, incluidos sus elementos ornamentales y decorativos, garantizando que el titular tenga el derecho exclusivo de fabricar, vender o importar productos con ese diseño.
- **Secretos comerciales:** Protegen la información confidencial que tiene valor comercial, como fórmulas, métodos, procesos o información técnica, impidiendo su divulgación no autorizada o su uso por parte de terceros.

4.2.2. Ley de propiedad industrial de Guatemala

Contexto histórico

La propiedad industrial en Guatemala se remonta al Convenio de París de 1883, del cual Guatemala fue signatario. Sin embargo, pocos años después, el país renunció al convenio debido a la carga financiera que representaba y la falta de solicitudes de registro de marcas. En 1886, la Asamblea Legislativa creó la primera Oficina de Patentes, dependencia del Ministerio de Fomento, anteriormente denominada como "Sección de Industrias". Posteriormente, en 1924, se creó la Oficina de Marcas y Patentes bajo el Decreto 882 (Registro de la Propiedad Intelectual de Guatemala, 2000).

A lo largo de los años, la administración y regulación de la propiedad industrial en Guatemala ha pasado por varios cambios organizativos. En 1944, la oficina se integró al Ministerio de Economía y Trabajo, y en 1956 se desligó para convertirse en una dependencia exclusiva del Ministerio de Economía. Desde 1975, Guatemala ha aplicado el Convenio Centroamericano, adoptando el nombre de Registro de la Propiedad Intelectual", según lo establecido en el artículo 164 de dicho convenio.

En 1998, se produjo una reestructuración importante cuando la institución pasó a denominarse Registro de la Propiedad Intelectual, abarcando tanto la Propiedad Industrial (marcas, patentes, diseños industriales, modelos de utilidad) como los Derechos de Autor y Derechos Conexos. Esta reorganización se formalizó con el Acuerdo Gubernativo 182-2000, que estableció al Registro de la Propiedad Intelectual como una dependencia del Ministerio de Economía.

Disposiciones generales

La ley de propiedad industrial establece el marco normativo para proteger y promover la crea-

tividad intelectual en el ámbito industrial y comercial. Enfocándose en aspectos específicos como la adquisición, mantenimiento y protección de signos distintivos, patentes de invención, modelos de utilidad y diseños industriales, así como la salvaguarda de secretos empresariales y medidas contra la competencia desleal.

Estipula que todas las personas, independientemente de su nacionalidad, domicilio o actividad, tienen derecho a beneficiarse de las protecciones otorgadas por esta ley. Se garantiza el principio de trato nacional, asegurando que las personas de otros Estados vinculados a Guatemala por tratados de trato nacional, así como aquellas con domicilio o establecimiento en Guatemala, puedan acceder a los mismos derechos que los ciudadanos guatemaltecos.

Estructura

La ley se compone de tres estructuras principales: títulos, capítulos y secciones, que organizan de manera sistemática el contenido. Esta organización sigue un enfoque introspectivo, permitiendo una exploración detallada de cada aspecto relevante de la regulación de la propiedad industrial en Guatemala. Esta orientación hacia el análisis interno facilita la comprensión y consulta de la ley, al proporcionar una estructura clara y coherente para su estudio y aplicación. La ley de propiedad industrial se compone de la siguiente forma:

- Título I: Normas Comunes
  - Capítulo Único: Disposiciones Generales
- Título II: De las marcas y otros signos distintivos
  - Capítulo I: De las marcas
  - Capítulo II: Marcas colectivas
  - Capítulo III: Marcas de certificación
  - Capítulo IV: Extinción del Registro de la Marca
  - Capítulo V: Expresiones o señales de Publicidad
  - Capítulo VI: Nombres comerciales
  - Capítulo VII: Emblemas
  - Capítulo VIII: Indicaciones Geográficas y Denominaciones de Origen
- Título III: Invenciones, modelos de utilidad y diseños industriales
  - Capítulo I: Invenciones
  - Capítulo II: Modelos de utilidad
  - Capítulo III: Diseños industriales
- Título IV: Del Registro de la Propiedad Intelectual
  - Capítulo I: Registro y Publicidad
  - Capítulo II: Clasificaciones

    - Capítulo III: Tasas y otros pagos
- Título V: De la Represión de la Competencia Desleal
    - Capítulo Único: Actos de Competencia Desleal
- Título VI: Acciones procesales
    - Capítulo I: Disposiciones generales
    - Capítulo II: Acciones civiles
    - Capítulo III: Acciones penales
- Título VII: Disposiciones transitorias y finales
    - Capítulo I: Disposiciones transitorias
    - Capítulo II: Disposiciones finales

4.3. Exploración de conceptos tecnológicos en la creación de sistemas interactivos

4.3.1. Interfaz de programación de aplicaciones (API)

Un API, o Interfaz de Programación de Aplicaciones, funciona como un puente de comunicación entre dos componentes de software. Este mecanismo permite que distintas aplicaciones se conecten entre sí de manera estructurada y predefinida, facilitando el intercambio de datos y la ejecución de acciones específicas.

En su funcionamiento, el API define un conjunto de reglas, protocolos y definiciones que especifican cómo las aplicaciones pueden solicitar y compartir información entre sí. Por lo tanto, actúa como un intermediario que estandariza la comunicación y permite que los programas interactúen de manera coherente y eficiente.

4.3.2. Tunelización segura

Los túneles seguros son una técnica empleada en redes informáticas para establecer conexiones protegidas entre dispositivos o redes a través de internet. Esta conexión encriptada garantiza la seguridad de los datos transmitidos, evitando posibles interceptaciones o ataques maliciosos (IBM, 2022).

Su aplicación es variada, siendo comúnmente utilizados en entornos empresariales para conectar sucursales remotas a una red central de manera segura. Además, son esenciales en servicios de acceso remoto como VPN, que permiten a los usuarios acceder a recursos privados desde ubicaciones externas mediante una conexión cifrada.

En el ámbito del desarrollo de software, los túneles seguros son utilizados para exponer temporalmente servicios y aplicaciones locales a internet de forma segura, facilitando pruebas y demostraciones sin necesidad de desplegar la aplicación en un servidor externo. Esto permite a los desarrolladores compartir fácilmente su trabajo y realizar demostraciones de funcionalidades.

4.3.3. *Microframework web*

Un *microframework web* es una herramienta ligera y minimalista diseñada para facilitar el desa-

rrollo de aplicaciones web de manera simple y eficiente. A diferencia de los *frameworks web* completos, los *microframeworks web* suelen ofrecer solo las funcionalidades esenciales necesarias para construir aplicaciones web básicas, lo que los hace más flexibles y fáciles de entender (Dhruv P., 2022).

Están diseñados para ser altamente modularizados y permiten a los desarrolladores agregar solo las características que necesitan, sin cargar con funcionalidades adicionales que podrían no ser utilizadas. Esto los hace especialmente adecuados para proyectos pequeños o prototipos rápidos donde la simplicidad y la velocidad de desarrollo son prioritarias.

4.3.4. *Prompting*

En el ámbito del procesamiento del lenguaje natural, *prompting* se refiere a la práctica de proporcionar una entrada específica o un contexto inicial para guiar la generación de texto por parte de un modelo. Esta técnica se utiliza para influir en el contenido y el estilo del texto generado, dirigiendo la atención del modelo hacia un tema particular o un tipo específico de respuesta (Sirit A., 2023).

En lugar de dejar que el modelo genere texto de forma completamente autónoma, el *prompting* establece un marco inicial que orienta la producción de texto hacia un resultado deseado. Esto puede implicar proporcionar una frase inicial que defina el contexto o el tema de la generación de texto subsiguiente, lo que ayuda al modelo a producir resultados más coherentes y relevantes para la tarea en cuestión.

5. Metodología

La metodología de este proyecto se estructura en cinco etapas clave, cada una destinada a abordar aspectos específicos del proceso de implementación y evaluación de un LLM *fine-tuning* supervisado para el análisis de la ley de propiedad intelectual. Estas etapas van desde la investigación y definición del alcance legislativo hasta el desarrollo de un sistema interactivo para gestionar consultas legales.

1. Investigación
2. Definición del alcance legislativo y recopilación de datos
3. Selección del modelo LLM y preparación de datos
4. Implementación del *fine-tuning* y evaluación del modelo
5. Consideraciones éticas y desarrollo del sistema de gestión de consultas

5.1. Investigación

La primera etapa del proyecto consiste en el desarrollo de una exhaustiva investigación en torno a los LLMs, la técnica *fine-tuning* y el estado del arte de la aplicación de *fine-tuning* en LLMs para su uso en el ámbito legal. Se llevaron a cabo revisiones bibliográficas y análisis de estudios previos para comprender las prácticas actuales, los desafíos y las tendencias emergentes en este campo. Este proceso proporcionó una base sólida para identificar las mejores prácticas, las herramientas y las técnicas más relevantes para el *fine-tuning* de LLMs en contextos legales específicos.

5.2. Definición del alcance legislativo y recopilación de datos

5.2.1. Alcance legislativo

En colaboración con el profesional legal Lic. Giancarlo Figueroa, con experiencia en tecnología y abogacía, se definió el alcance legislativo del proyecto. Tras considerar diversas opciones, se optó por la ley de propiedad intelectual, específicamente la ley de propiedad industrial. Esta elección se basó en varios criterios, entre ellos, la actualización y disponibilidad virtual de la ley, así como su longitud, que resultaba adecuada para los límites de tiempo del proyecto. Se discutió la posibilidad de incluir jurisprudencia, sin embargo, las limitaciones tecnológicas, como la falta de virtualización de documentos, llevaron al asesor técnico MSc. Alberto Suriano, ingeniero en ciencias computacionales y experto en inteligencia artificial, a recomendar su exclusión debido a las restricciones de tiempo establecidas.

5.2.2. Recopilación de datos

Para la recolección de datos, se adoptaron las recomendaciones discutidas durante la selección del alcance. En el contexto de la ley de propiedad industrial, se determinó que el texto legal requerido estaba disponible en el portal oficial del Registro de Propiedad Intelectual de Guatemala. La elección de esta fuente se justificó debido a su accesibilidad en formato PDF a través de la plataforma mencionada, lo que evitó la necesidad de destinar recursos adicionales para la búsqueda de otras fuentes. Además, se evaluó que este formato cumplía con los requisitos necesarios para su posterior procesamiento. Es fundamental resaltar que se ejerció un cuidado especial para asegurar que el ma-

terial recopilado no constituyera una versión comentada o explicada de la ley, con el propósito de evitar cualquier conflicto potencial con los autores y salvaguardar la integridad de los datos utilizados en el proyecto.

5.3. Selección del LLM y preparación de datos

5.3.1. Selección del LLM

Dada la diversidad de modelos disponibles para diversos propósitos, se volvió imperativo delimitar el alcance del proyecto en relación con la cantidad de observaciones de entrenamiento mínimas requeridas para alcanzar el objetivo. Considerando factores previamente discutidos, como la longitud de la ley de propiedad industrial, el tiempo de ejecución del proyecto, la posible composición de las observaciones (ya sea casos o interrogantes comunes en relación a la ley seleccionada) y el proceso de obtención de las mismas (manual o semi-manual), se concluyó que el conjunto de datos necesario para llevar a cabo el proceso de *fine-tuning* sería probablemente reducido, oscilando entre 1000 y 5000 observaciones. Por consiguiente, se hizo necesario optar por modelos ampliamente reconocidos en la comunidad debido a su eficacia en el proceso de *fine-tuning* con conjuntos de datos de pequeñas dimensiones. En base a esto se escogieron los siguientes modelos: GPT-3.5-turbo, BERT y T5-large, los cuales serían sometidos a comparación bajo los siguientes criterios:

1. **Capacidad y tamaño del modelo**

Cuadro 1: Capacidad y tamaño de los modelos propuestos

Modelo	Parámetros	Tamaño (memoria)
gpt-3.5-turbo	20B	N/A*
BERT	340M	1.2GB
T5-large	770M	2.95GB

NOTA: gpt-3.5-turbo no aplica (N/A) dado que ya cuenta con una infraestructura con OpenAI.

2. **Arquitectura del modelo**

Cuadro 2: Arquitectura y potencial de arquitectura de los modelos propuestos

Modelo	Arquitectura	Potencial de arquitectura
gpt-3.5-turbo	Transformer unidireccional	Generación de texto coherente
BERT	Transformer bidireccional	Clasificación y extracción de información
T5-large	Transformer text-to-text	Traducción y resúmenes

3. **Pre-entrenamiento multilingüe**

   Los tres modelos seleccionados demuestran competencia en la comprensión y ejecución de instrucciones en español, habiendo sido pre-entrenados utilizando conjuntos de datos en dicho idioma. No obstante, es importante destacar que su rendimiento está intrínsecamente ligado a la magnitud de sus respectivos parámetros.

4. **Disponibilidad de recursos**

   Tanto BERT como T5-base son de código abierto, lo que significa que no vienen con una infraestructura de ejecución integrada (Toolify.ai, 2024). Por lo tanto, se requiere disponer de recursos adecuados para alojar el modelo, realizar el proceso de fine-tuning y acceder eficientemente a los resultados. En caso de no contar con la infraestructura necesaria, es posible recurrir a servicios de alojamiento virtualizados, lo que implica costos por hora en función de

las capacidades requeridas y los gastos asociados a las consultas.

Por otro lado, GPT-3.5-turbo de OpenAI ofrece una infraestructura preparada para llevar a cabo estos trabajos, con tarifas basadas únicamente en el número de tokens entrenados y las consultas realizadas.

5. **Uso y disponibilidad**

   Aunque las tres opciones disponen de una amplia variedad de fuentes de información para ejecutar el *fine-tuning* de manera eficiente, en términos de facilidad de ejecución, la implementación de BERT y T5-large requiere un esfuerzo adicional en términos de capacitación técnica comparado con el modelo GPT-3.5-turbo, debido a la necesidad de establecer una infraestructura propia para la ejecución. Esto implica la adquisición virtual o configuración física del equipo necesario para alojar el modelo, llevar a cabo el proceso de *fine-tuning* y gestionar las consultas a los resultados.

   Por el contrario, al aprovechar la infraestructura proporcionada por OpenAI en el modelo GPT-3.5-turbo, se disminuye significativamente la necesidad de invertir tiempo en configurar completamente el entorno para entrenar un modelo de lenguaje de gran escala. Esto permite a los usuarios concentrarse directamente en el proceso de *fine-tuning*, calidad de datos y en el análisis de los resultados, sin tener que preocuparse por la configuración o gestión de la infraestructura.

Basándonos en los criterios previamente establecidos, GPT-3.5-turbo se perfila como la elección más adecuada para el proyecto de fine-tuning destinado a especializar el modelo en la ley de propiedad industrial/intelectual en Guatemala. En contraste con las alternativas evaluadas, BERT y T5-large, GPT-3.5-turbo destaca por su impresionante capacidad de 20 mil millones de parámetros (20B), situándolo como un modelo de gran envergadura y potencial. Además, su arquitectura basada en *Transformer* unidireccional lo hace especialmente apto para generar texto coherente, una habilidad esencial para abordar las sutilezas y complejidades inherentes a la ley de propiedad industrial/intelectual.

En términos de disponibilidad de recursos, GPT-3.5-turbo se beneficia de una infraestructura lista para su uso gracias a su asociación con OpenAI, eliminando así la necesidad de adquirir o configurar equipo adicional. Este aspecto simplifica considerablemente la ejecución del proyecto, permitiendo a los usuarios concentrarse directamente en el proceso de *fine-tuning*, la mejora de la calidad de los datos y el análisis de los resultados. En resumen, GPT-3.5-turbo ofrece una combinación única de capacidad, arquitectura adecuada y disponibilidad de recursos, lo que lo convierte en la opción preferida para llevar a cabo el proyecto de especialización en la ley de propiedad industrial/intelectual en Guatemala.

5.3.2. Preparación de datos

Para llevar a cabo la preparación de los datos que serán utilizados en el modelo, es fundamental considerar el tipo de *fine-tuning* que se llevará a cabo y seleccionar el formato final de las observaciones en consecuencia. En este sentido, se propusieron inicialmente dos alternativas:

1. **Aprendizaje por refuerzo a partir de la retroalimentación humana (RLHF):** Durante la etapa de planificación del proyecto, se confirmó la participación de expertos legales con el propósito de brindar orientación y asegurar la adopción de un enfoque completo y realista. La naturaleza de este método, basado en la retroalimentación humana, sugería que su implementación sería efectiva gracias a la intervención de los asesores legales implicados en el proyecto.

   Sin embargo, la aplicación de este método conllevaba una serie de consideraciones económicas importantes. Esto incluía el pago por las horas de asesoramiento de los expertos legales, quienes

colaborarían en la creación, validación, retroalimentación y corrección tanto de los insumos como de los resultados. Asimismo, se debían tener en cuenta los costos asociados con la ejecución de múltiples versiones del proyecto para validar la opción más óptima.

2. ***Fine-tuning* supervisado:** Tras discutir con el asesor MSc. Alberto Suriano, se destacó que este enfoque es más viable económicamente y ofrece la ventaja de evaluar de manera más objetiva los resultados del fine-tuning a través de métricas como la exactitud, la precisión y la recuperación. Aunque el proceso siempre requeriría la participación de expertos legales para certificar la calidad de los datos ingresados, el tiempo dedicado a su asesoramiento sería considerablemente menor en comparación con el aprendizaje por refuerzo.

En base a las consideraciones anteriores, se decidió utilizar el método de *fine-tuning* supervisado, justificándose por las siguientes razones fundamentales. En primer lugar, este enfoque ofrece una mayor capacidad para evaluar y comprender de manera objetiva los resultados obtenidos. Al utilizar métricas estándar como la exactitud, la precisión y la recuperación, se puede medir de manera más clara y precisa el rendimiento del modelo en comparación con el aprendizaje por refuerzo, donde la evaluación puede ser más subjetiva y compleja.

Además, el fine-tuning supervisado presenta una mayor eficiencia en términos de tiempo y recursos. Aunque aún se requiere la participación de expertos legales para certificar la calidad de los datos ingresados, el tiempo dedicado a este asesoramiento es generalmente menor en comparación con el aprendizaje por refuerzo. Esto se debe a que el proceso de fine-tuning supervisado se basa en un conjunto predefinido de datos de entrenamiento, lo que permite una implementación más rápida y menos dependiente de la retroalimentación constante de los expertos (Khawaja, R., 2023).

Segmentación de la ley de propiedad industrial de Guatemala

Dada la elección del método de *fine-tuning* supervisado, el proceso de segmentación de la ley de propiedad industrial involucra la definición de etiquetas y variables que explican la misma.

Considerando el propósito del proyecto (proporcionar orientación legal en consultas específicas), se propuso que las etiquetas fueran títulos creados a partir de fragmentos de ley representativos y útiles, estos títulos deberían ser capaces de describir el rango del fragmento como la esencial del mismo. Por otra parte, las observaciones que describirían las etiquetas se conformarían por simulaciones de casos cotidianos/comunes o preguntas relevantes con respecto a la ley contenida en el fragmento.

Es importante considerar que la ley de propiedad industrial tiene una estructura que varía desde títulos hasta artículos individuales, lo que requiere un enfoque equilibrado en la segmentación para evitar afectar negativamente el desempeño del *fine-tuning*.

Creación de etiquetas

Para determinar el número y tamaño de los fragmentos, se analizó el número de páginas por título, observando una variabilidad significativa en algunos títulos en comparación con otros. Se calculó un promedio de páginas por título, excluyendo aquellos con variaciones atípicas, y se determinó un tamaño promedio de fragmento. Sin embargo, este enfoque provocaba una perdida de contexto entre fragmentos debido a la intersección entre títulos, capítulos y secciones, por lo que tuvo que ajustarse manualmente para evitar esta pérdida y asegurar que fueran significativos. Como resultado, se estableció un total de 18 fragmentos correspondientes a las etiquetas del *fine-tuning* supervisado, asegurando una representación coherente y significativa de la ley en su conjunto.

Una vez delimitados y ajustados los rangos de los fragmentos, se procedió a enriquecerlos con una breve descripción contextual, siguiendo la propuesta establecida. Con el propósito de alcanzar este objetivo, se emplearon tres enfoques complementarios:

1. **Análisis de la esencia mediante *prompting*:** Se recurrió al modelo GPT-3.5 turbo para obtener una comprensión más profunda del contenido de cada fragmento, lo que permitió capturar su esencia de manera detallada y precisa.

2. **Lectura directa de los fragmentos:** Se llevó a cabo una revisión manual de los fragmentos para asegurar la coherencia entre el análisis del modelo y el contenido real de la ley, garantizando así la fidelidad de las descripciones agregadas.

3. **Consulta y validación con un experto en leyes de propiedad intelectual:** Se contó con la asesoría del Lic. Edwin Menchú, un experto en comercio internacional, quien corroboró y validó el contexto añadido a cada etiqueta, asegurando su pertinencia y precisión en el ámbito legal.

Como resultado de este proceso, se generaron las siguientes etiquetas finales, las cuales encapsulan tanto el rango del fragmento como los distintos aspectos esenciales del mismo:

1. TÍTULO 2 CAPÍTULO 1 SECCIÓN 1: Marcas: normas, adquisición, prioridad, derechos, inadmisibilidad en Guatemala

2. TÍTULO 2 CAPÍTULO 1 SECCIÓN 2: Registro de marcas: requisitos, procedimientos y efectos legales específicos.

3. TÍTULO 2 CAPÍTULO 1 SECCIÓN 3-5: Renovación, corrección, limitación, derechos y licencia en marcas registradas.

4. TÍTULO 2 CAPÍTULO 2 NO SECCIÓN: Normas aplicables, solicitud, reglamento, examen, registro, cambios, licencia, uso.

5. TÍTULO 2 CAPÍTULO 3 NO SECCIÓN: Normas, titularidad, registro, vigencia, uso, gravamen, enajenación, reserva, extinguida.

6. TÍTULO 2 CAPÍTULO 4-6 NO SECCIÓN: Extinción: vencimiento, caducidad, cancelación, generización, falta de uso, sentencia.

7. TÍTULO 2 CAPÍTULO 8 SECCIÓN 1 y 2: Protección de indicaciones geográficas y denominaciones de origen en Guatemala.

8. TÍTULO 3 CAPÍTULO 1 SECCIÓN 1: Protección legal para invenciones; requisitos y derechos del inventor.

9. TÍTULO 3 CAPÍTULO 1 SECCIÓN 2: Trámite, prioridad, examen, requisitos, solicitud, invención, descripción, publicación, resolución, certificado.

10. TÍTULO 3 CAPÍTULO 1 SECCIÓN 3 y 4: División, modificación, conversión, corrección, enajenación, vigencia, ajuste, protección, limitaciones, agotamiento.

11. TÍTULO 3 CAPÍTULO 1 SECCIÓN 5-7: Licencias contractuales y obligatorias para explotar patentes según disposiciones legales.

12. TÍTULO 4 CAPÍTULO 1-3 NO SECCIÓN: Registro público de propiedad intelectual; tasas y clasificaciones definidas.

13. TÍTULO 5 CAPÍTULO ÚNICO NO SECCIÓN: Represión de competencia desleal: actos prohibidos y protección de secretos.

14. TÍTULO 6 CAPÍTULO 1 NO SECCIÓN: Protección judicial y medidas contra la competencia desleal.

15. TÍTULO 6 CAPÍTULO 2 SECCIÓN 1 y 2: Procedimientos legales para proteger derechos industriales y combatir competencia desleal en Guatemala.

16. TÍTULO 6 CAPÍTULO 2 SECCIÓN 3-5: Acciones legales para proteger derechos industriales y combatir competencia desleal en Guatemala.

17. TÍTULO 6 CAPÍTULO 3 NO SECCIÓN: Acciones penales y medidas cautelares en casos de infracciones a la Propiedad Industrial en Guatemala

18. TÍTULO 7 CAPÍTULO 1 y 2 NO SECCIÓN: Disposiciones transitorias y finales para la implementación de la Ley de Propiedad Industrial en Guatemala.

Creación de observaciones

Una vez identificadas las etiquetas, se procedió a generar las observaciones que describirían estos fragmentos y servirían en el futuro como casos de entrenamiento para el modelo. Según lo propuesto, estas observaciones serían casos cotidianos o comunes, referentes a la ley de propiedad industrial que un profesional podría experimentar día a día. Se contó con la colaboración del asesor legal Lic. Edwin Menchú, quien proporcionó ejemplos basados en su experiencia práctica para orientar esta tarea.

A partir de esto se formuló una estructura generalizada consciente de la diversidad de contextos y fragmentos de ley, que funcionaría como base al momento de desarrollar los casos. Esta estructura incluyó:

1. Introducción del caso

2. Descripción de la acción legal

3. Autoridades o entidades involucradas

4. Procedimiento legal

5. Fundamento legal

6. Resolución o resultado

Se generaron los primeros casos en colaboración con el asesor legal para establecer un marco de referencia. Posteriormente, se acordó revisar muestras aleatorias una vez se alcanzara el cincuenta por ciento del conjunto final de datos de entrenamiento.

Para estimar el número adecuado de observaciones, se consultó la documentación de OpenAI (OpenAI, 2023a), donde se recomendaba ejecutar la primera iteración de *fine-tuning* con cincuenta casos por etiqueta. En base a esta información, se determinó un número de novecientos casos para el conjunto de datos de entrenamiento.

Debido a la complejidad y extensión del proceso de generación de casos, se estableció la realización de revisiones muestrales aleatorias una vez se hubiera alcanzado el treinta por ciento del conjunto de datos. Posteriormente, al haber generado esta fracción inicial de casos, se procedió a contactar nuevamente al asesor para verificar la estructura y recibir retroalimentación.

Una vez confirmada la calidad del conjunto de datos por el asesor, se continuó con la generación de los casos restantes mediante el modelo GPT-3.5. Para ello, se utilizó como entrada la estructura

previamente definida para la creación de casos, junto con ejemplos de la misma y el fragmento de ley correspondiente. Durante este proceso, se puso un énfasis especial en el uso del *prompting*. Esto garantizó que las observaciones generadas mantuvieran coherencia con la estructura establecida y reflejaran de manera precisa el contexto legal y las directrices proporcionadas por el asesor.

Una vez completado el conjunto de entrenamiento, se programó una sesión adicional con el asesor para llevar a cabo una verificación aleatoria final. Durante esta etapa, se revisaron aleatoriamente una selección de casos para garantizar la coherencia y calidad del conjunto de datos en su totalidad. Posteriormente, se acordó iniciar la creación del conjunto de prueba, que constaría de diez casos por etiqueta. Este conjunto de prueba serviría para evaluar el desempeño del modelo y garantizar su capacidad para proporcionar respuestas exactas y contextualizadas en una variedad de situaciones legales.

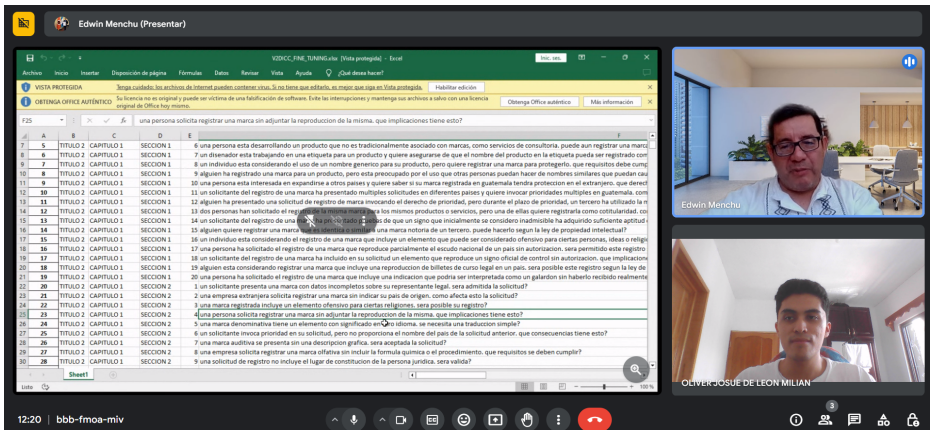


Figura 1: Reunión con el asesor legal Lic. Edwin Menchú

Normalización del conjunto de datos

Después de la generación de etiquetas y casos, se llevó a cabo un proceso de normalización con el objetivo de asegurar la coherencia y la consistencia en los datos. Esta fase implicó la eliminación de mayúsculas, acentos y cualquier símbolo que careciera de relevancia para el contexto legal.

La normalización de etiquetas y casos desempeña un papel crucial en la comprensión adecuada del texto por parte del modelo de lenguaje, permitiéndole generalizar de manera efectiva. Al eliminar las variaciones innecesarias, como las diferencias de capitalización o los signos diacríticos, se simplifica la estructura del texto, facilitando su procesamiento por parte del modelo.

Este procedimiento no solo contribuye a la coherencia y uniformidad de los datos, sino que también ayuda a prevenir inconsistencias y posibles errores durante el entrenamiento y la evaluación del modelo. Al estandarizar el formato del texto, se promueve una mayor calidad en el desempeño del modelo en diversas tareas relacionadas con la comprensión y la generación de lenguaje natural.

Formateo del conjunto de datos

Una vez completada la generación y normalización de toda la base de datos, el siguiente paso fue dividir el conjunto en dos grupos: uno de entrenamiento, que representaba el ochenta y tres por ciento del total (correspondiente a cincuenta observaciones por etiqueta), y otro de prueba, que comprendía el diesisiete por ciento restante (correspondiente a diez observaciones por etiqueta).

Dada la selección del LLM, se consultó nuevamente la documentación de OpenAI para determinar el formato adecuado de ingreso para ejecutar el *fine-tuning* (OpenAI, 2023a). Siguiendo las

pautas establecidas en los estándares, se identificó la necesidad de utilizar el formato JSON para cumplir esta tarea. En base a la selección del *fine-tuning* supervisado, las necesidades del proyecto y los requerimientos de la entrada del modelo, se propuso la siguiente estructura:

```
{
    'messages':[
        {'role':'system','content':system_prompt},
        {'role':'user','content':case},
        {'role':'assistant','content':label}
    ]
}
```

Figura 2: Estructura JSON propuesta para desarrollo del *fine-tuning* supervisado

Dentro de esta estructura, se definen tres roles relevantes:

1. **El rol '*system*':** Representa el sistema o el *prompt*, utilizado para solicitar al modelo que clasifique un caso en alguna de las etiquetas creadas, y posteriormente justifique su respuesta.

2. **El rol '*user*':** Representa al futuro usuario ingresando un caso cotidiano/común.

3. **El rol '*assistant*':** Representa la respuesta que el modelo predecirá, indicando la etiqueta o fragmento de ley al que pertenece el caso con su respectiva justificación.

Una vez definida la estructura del JSON según los roles de sistema, usuario y asistente, se procedió a convertir los conjuntos de datos en archivos JSON listos para su procesamiento en el modelo de *fine-tuning*.

Cada conjunto de datos, tanto el de entrenamiento como el de prueba, se transformó en un archivo JSON que sigue la estructura especificada, asegurando así la coherencia y consistencia en el formato de entrada requerido por el modelo de lenguaje.

5.4. Implementación del *fine-tuning* y evaluación del modelo

5.4.1. Implementación del *fine-tuning*

Para implementar el proceso de *fine-tuning* del modelo GPT-3.5, primero fue necesario adquirir acceso al API de OpenAI mediante la creación de una cuenta y la obtención del token correspondiente a través de la interfaz de acceso proporcionada. Una vez obtenido el token, se procedió a utilizar la biblioteca de OpenAI para Python para interactuar con el API y realizar pruebas de conexión preliminares.

Posteriormente, se utilizó la biblioteca para cargar los datos de entrenamiento y prueba en formato JSON según la estructura acordada, especificando roles de sistema, usuario y asistente. Estos archivos se cargaron en el API como conjuntos de entrenamiento y prueba respectivamente.

Una vez realizadas las cargas, era necesario iniciar un trabajo de *fine-tuning* definiendo el modelo base a utilizar y configurando hiperparámetros básicos como el número de épocas.

En una primera instancia, se llevó a cabo un trabajo inicial utilizando el modelo GPT-3.5-turbo-

0125 con una sola época, siguiendo las recomendaciones proporcionadas en la documentación de OpenAI para las primeras iteraciones del *fine-tuning* (OpenAI, 2023a). Sin embargo, se propuso realizar una segunda instancia con dos épocas con el objetivo de comparar el rendimiento entre estas dos configuraciones y determinar la necesidad de ajustar este parámetro en futuras iteraciones del proyecto. Es importante destacar que no se exploraron trabajos con un número mayor de épocas debido a las consideraciones económicas asociadas a este hiperparámetro. Una vez completadas ambas ejecuciones, se monitoreó el progreso de los trabajos hasta su finalización, momento en el cual los modelos ajustados (*fine-tuned*) estuvieron listos para la evaluación.

Figura 3: Finalización del proceso de fine-tuning

5.4.2. Evaluación del modelo

Para evaluar el modelo, se consideraron cuatro factores clave y se llevaron a cabo procedimientos específicos:

1. **Pérdida en la validación completa (*Full Validation Loss*):** Este valor, generado por OpenAI al culminar el proceso de entrenamiento, refleja el nivel de error del modelo en un conjunto de datos de validación independiente. Su seguimiento es esencial en el contexto del *fine-tuning*, ya que permite comparar el rendimiento entre modelos y guiar la elección de hiperparámetros. Esta evaluación contribuye de manera significativa a mejorar la capacidad de generalización del modelo y su efectividad en tareas específicas. A partir de esta métrica, se llevaron a cabo comparaciones entre las dos variantes de *fine-tuning* propuestas, con 1 y 2 épocas, respectivamente.

2. **Exactitud (*Accuracy*):** Esta métrica constituye otro aspecto crucial, brindando una visión sobre la exactitud de las predicciones o recomendaciones del modelo. Se llevó a cabo una evaluación detallada de esta métrica tanto en el modelo base, gpt-3.5-turbo-0125, como en el modelo ajustado (*fine-tuned*), considerando cada una de sus variantes de hiperparámetros (1 y 2 épocas). Este enfoque permitió comparar exhaustivamente los resultados obtenidos, identificando la mejora que el proceso de *fine-tuning* aporta a una tarea específica y la importancia de evaluar variaciones en los hiperparámetros.

3. **Precisión (*Precision*):** Esta métrica es esencial para evaluar la efectividad del modelo en la identificación precisa de casos positivos dentro de un conjunto de datos. Se realizó un análisis de precisión en el modelo base y las dos variantes del modelo ajustado (*fine-tuned*), considerando 1 y 2 épocas de entrenamiento respectivamente. Esto permitió comparar su desempeño y determinar si la implementación del *fine-tuning* y el aumento en el número de épocas podría ser necesario para mejorar el rendimiento del modelo en este aspecto.

4. **Recuperación (*Recall*):** Esta métrica es fundamental para evaluar la proporción de casos positivos que fueron correctamente identificados por el modelo entre todos los casos positivos reales. Se realizó una evaluación de esta métrica en el modelo base como en las dos variantes disponibles del modelo ajustado (*fine-tuned*) (1 y 2 épocas), con el objetivo de comparar su desempeño y determinar si existe una mejora al implementar *fine-tuning* o al aumentar el número de épocas.

Una vez obtenidas las métricas, se procedió a llevar el modelo ajustado (*fine-tuned*) a pruebas reales con la colaboración del asesor legal Lic. Edwin Menchú. Esta etapa de pruebas en un entorno real proporciona una validación adicional del desempeño del modelo y su capacidad para abordar situaciones del mundo real en el ámbito legal.

5.5. Consideraciones éticas y desarrollo del sistema de gestión de consultas

5.5.1. Consideraciones éticas

El proyecto aborda varias consideraciones éticas fundamentales:

1. **Respeto a la propiedad intelectual:** La decisión de evitar el uso de leyes comentadas o explicadas ayuda a evitar problemas relacionados con la propiedad intelectual de los autores originales. Esto garantiza que se respeten los derechos de autor y se eviten posibles disputas legales.

2. **Imparcialidad y sesgo:** La elección de no incluir comentarios de autores externos al organismo creador de la ley contribuye a mantener la imparcialidad en el sistema judicial. Evitar la introducción de sesgos externos es crucial para garantizar que la justicia sea impartida de manera equitativa y objetiva.

3. **Protección de la privacidad:** El uso de casos genéricos en lugar de entidades o individuos reales protege la privacidad y confidencialidad de las personas involucradas. Esto es esencial para mantener la integridad del proyecto y proteger los derechos y la privacidad de los individuos.

4. **Confidencialidad y responsabilidad:** Mantener en secreto los casos y la documentación pertinente utilizada en el proyecto, así como buscar la confirmación de calidad por parte del asesor legal, demuestra un alto nivel de responsabilidad y compromiso ético. Esto garantiza que se manejen los datos de manera confidencial y se trabaje con estándares éticos elevados.

5. **Transparencia y divulgación:** Explicar claramente la estructura y el proceso utilizado en el proyecto a través de un trabajo escrito muestra transparencia y claridad en la metodología empleada. Esto ayuda a generar confianza en el proyecto y a demostrar el compromiso con la integridad y la ética en la investigación.

6. **Beneficio social:** El propósito final del proyecto, que es mejorar la experiencia de las consultas legales para los profesionales del derecho y aquellos que necesitan ayuda legal, refleja un objetivo ético de contribuir al bienestar de la sociedad. Al proporcionar información de calidad

y mejorar el acceso a la justicia, el proyecto busca beneficiar a un amplio espectro de personas y comunidades.

5.5.2. Desarrollo del sistema de gestión de consultas

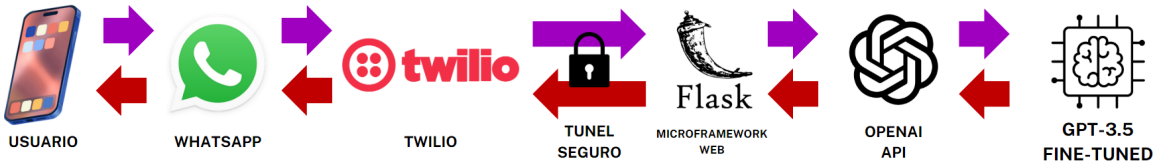


Figura 4: Flujo de ejecución sistema de gestión de consultas

El desarrollo del sistema de gestión de consultas constituye un proceso meticuloso y estratégico que culmina en la disponibilidad pública del modelo ajustado (*fine-tuned*) a través de una interfaz accesible y amigable para los usuarios. La elección de WhatsApp como plataforma de comunicación se fundamenta en su amplia penetración en la sociedad guatemalteca, alcanzando aproximadamente el cuarenta y cinco por ciento de la población para el año 2023, respaldada por múltiples compañías de comunicación nacional que ofrecen servicios especializados en esta aplicación.

El flujo de interacción se concibe como la interceptación de mensajes del usuario mediante WhatsApp, los cuales son dirigidos a un servidor intermedio para establecer la conexión con el modelo ajustado (*fine-tuned*) y, posteriormente, enviar la respuesta al usuario inicial. Para lograr esta conexión inicial, se optó por la utilización de Twilio, que proporciona una infraestructura ágil y económicamente viable. La implementación de un entorno de pruebas con créditos de bienvenida contribuyó a reducir los costos totales del proyecto.

Una vez establecida la conexión con Twilio, se desarrolló una comunicación segura mediante un túnel protegido hacia el servidor intermedio. La creación de este último se llevó a cabo mediante un servidor local implementado con Flask, un *microframework web* idóneo para pruebas de concepto, como la publicación del modelo a través de WhatsApp. Posteriormente, se estableció la comunicación bidireccional con OpenAI para acceder al modelo ajustado (*fine-tuned*).

Es crucial destacar que los costos de procesamiento del modelo ajustado (*fine-tuned*) es tres veces superior a la del modelo base (gpt-3.5-turbo-0125), lo que generó la necesidad de incorporar medidas de filtrado y seguridad en el servidor local. Estas medidas abarcan desde la identificación de consultas válidas hasta el rechazo de mensajes fuera del ámbito previsto por el modelo, con el fin de evitar gastos innecesarios derivados de consultas inapropiadas o maliciosas.

Finalmente, se implementó un algoritmo de detección de indicios de ayuda para orientar a los usuarios sobre cómo interactuar y realizar consultas de manera efectiva. Este enfoque refleja el compromiso del proyecto con la mejora continua y la experiencia del usuario, garantizando una interacción segura y productiva con el sistema de gestión de consultas.

5.5.3. Validación del sistema de gestión de consultas

Para validar la efectividad del sistema de gestión de consultas, se evaluó en dos aspectos principales:

1. **Creación y envío de *prompts*:** Se implementó un proceso de interacción directa con el sistema con el objetivo de validar su correcto funcionamiento y asegurar la adherencia al flujo

propuesto. Este proceso implicó un monitoreo exhaustivo de cada etapa del flujo de mensajes a través de la infraestructura del sistema. Se verificó que los mensajes fueran interceptados adecuadamente, ingresaran al túnel seguro para preservar la privacidad de la información, se procesaran de manera eficiente en el servidor local basado en Flask y finalmente, se estableciera la interacción con el modelo ajustado (*fine-tuned*) para la generación de respuestas precisas.

2. **Guiamiento del usuario en la creación de *prompts* e interpretación de las respuestas:** Para asegurar que el sistema proporcionara orientación al usuario tanto en la creación como en la interpretación de las respuestas, se llevó a cabo una interacción directa con el sistema para evaluar la ejecución del algoritmo diseñado para detectar indicios de ayuda. Este algoritmo tenía como objetivo identificar si el usuario estaba teniendo problemas con la interacción y a partir de estos indicios buscaba orientar a los usuarios sobre cómo interactuar y realizar consultas de manera efectiva. Durante esta interacción, se verificó si el sistema ofrecía información detallada sobre cómo estructurar un *prompt* adecuado y qué significaba la salida proporcionada por el modelo ajustado (*fine-tuned*).

   Además, se confirmó que el sistema guiara al usuario en el proceso de creación del *prompt*, destacando su negativa a discutir sobre temas distintos a la propiedad intelectual.

6. Resultados

6.1. Pérdida de validación completa (*Full Validation Loss*)

Cuadro 3: Pérdida de validación completa por épocas

Modelo	*Full Validation Loss*
gpt-3.5-turbo-0125:9HlXk9TZ (1 época de *fine-tuning*)	0.1016
gpt-3.5-turbo-0125:9KV79yHU (2 épocas de *fine-tuning*)	0.1102

Los resultados de la pérdida de validación completa revelan diferencias relevantes entre los dos modelos ajustados (*fine-tuned*).

El modelo gpt-3.5-turbo-0125:9HlXk9TZ, ajustado durante una sola época, exhibe una pérdida de validación completa de 0.1016, lo que sugiere un buen rendimiento en la tarea específica. Este valor indica el nivel de error del modelo en un conjunto de datos de validación independiente, donde una pérdida menor se correlaciona con un mejor rendimiento. Por otro lado, el modelo gpt-3.5-turbo-0125:9KV79yHU, ajustado (*fine-tuning*) durante dos épocas, muestra una pérdida ligeramente mayor de 0.1102. Aunque la diferencia en las pérdidas entre ambos modelos es relativamente pequeña, no se puede concluir definitivamente que agregar una época adicional de *fine-tuning* no mejora significativamente el rendimiento del modelo en términos de pérdida de validación completa.

Es importante tener en cuenta que este hallazgo podría indicar una posible tendencia, pero no permite afirmar con certeza el tipo específico de comportamiento, ya que existen diversas formas de evolución de datos, como lineal, exponencial, logarítmica, polinómica, entre otras. Además, esta observación está limitada por la disponibilidad de fondos para realizar un *fine-tuning* prolongado. Por lo tanto, se requeriría una investigación más exhaustiva con una asignación de fondos más amplia para determinar con certeza si un cambio en las épocas de *fine-tuning* conlleva un impacto significativo en el rendimiento del modelo.

Estos resultados proporcionan una base sólida para tomar decisiones informadas sobre la optimización del rendimiento del modelo y la selección de hiperparámetros en proyectos futuros de *fine-tuning*.

6.2. Exactitud (*Accuracy*)

Cuadro 4: Exactitud por modelo

Modelo	Exactitud (*Accuracy*)
gpt-3.5-turbo-0125 (base)	5.00 %
DISC-LawLLM (*Patent Agent Examination*)	40.68 %
gpt-3.5-turbo-0125:9HlXk9TZ (1 época de *fine-tuning*)	69.44 %
gpt-3.5-turbo-0125:9KV79yHU (2 épocas de *fine-tuning*)	58.33 %

La evaluación de la exactitud (*Accuracy*) La evaluación revela patrones interesantes en el desempeño de los diversos modelos, especialmente al compararlos con un modelo base y un proyecto similar aplicado en un contexto diferente.

El modelo base, gpt-3.5-turbo-0125, exhibe una precisión muy baja del 5.00 %, lo que significa

que acierta la clasificación de 1 caso en 20. Esta baja exactitud sugiere un rendimiento insatisfactorio para la tarea en cuestión, lo que subraya la necesidad de realizar un *fine-tuning* para adaptar el modelo a la tarea específica

Por otro lado, el proyecto DISC-LawLLM, que se enfoca en un contexto legal similar pero específicamente en el examen de agente de patentes, logra una exactitud notablemente más alta del 40.68 % o aproximadamente 2 casos correctos de cada 5. Esta disparidad en el rendimiento entre el modelo base y el proyecto chino indica la importancia de adaptar los modelos de lenguaje natural a tareas especializadas dentro de un dominio específico, como la ley de propiedad industrial.

En cuanto a los modelos ajustados (*fine-tuned*), los resultados muestran que el modelo gpt-3.5-turbo-0125:9HlXk9TZ, ajustado durante una sola época, alcanza una exactitud del 69.44 %, prediciendo aproximadamente 7 de cada 10 casos, la más alta entre los modelos presentados. Esto sugiere que incluso un *fine-tuning* relativamente breve puede mejorar significativamente el rendimiento del modelo en la tarea específica de propiedad industrial.

Sin embargo, es interesante observar que el modelo gpt-3.5-turbo-0125:9KV79yHU, ajustado durante dos épocas, muestra una exactitud ligeramente más baja del 58.33 % en comparación con el modelo ajustado (*fine-tuned*) de una sola época. Esto puede deberse a una posible sobreajuste o a una falta de generalización del modelo después de un *fine-tuning* prolongado. A pesar de la diferencia de 11.11 puntos porcentuales en las exactitudes, no se puede concluir de manera definitiva que la inclusión de una época adicional de *fine-tuning* no vaya a mejorar el rendimiento de la exactitud. Este hallazgo sugiere una posible tendencia, aunque no proporciona una certeza sobre el tipo específico de comportamiento, dado que los datos pueden evolucionar de diversas maneras, como lineal, exponencial, logarítmica, polinómica, entre otras. Además, es importante considerar que esta observación se ve limitada por la disponibilidad de fondos para llevar a cabo un *fine-tuning* prolongado.

En sinergia, estos resultados destacan la importancia del *fine-tuning* para adaptar los modelos de lenguaje natural a tareas especializadas, como la ley de propiedad industrial. Además, subrayan la necesidad de asignar recursos adicionales para investigar de manera más detallada si un aumento en las épocas de *fine-tuning* tiene un impacto significativo en la exactitud del modelo.

6.3. Precisión (*Precision*)

La precisión en el contexto de este proyecto se refiere a la cantidad de casos que el modelo predijo correctamente en comparación con todas las predicciones realizadas sobre una etiqueta o fragmento de ley. El modelo base mostró una precisión del 11.11 %, lo que significa que solo 1 de cada 9 predicciones realizadas específicamente sobre una clase, fue correcta. Por otro lado, los modelos ajustados (*fine-tuned*) exhibieron resultados considerablemente mejores, lo que subraya la importancia de implementar procesos de *fine-tuning* para mejorar el rendimiento en tareas especializadas.

Cuadro 5: Precisión por modelo

Modelo	Precisión (*Precision*)
gpt-3.5-turbo-0125 (base)	11.11 %
gpt-3.5-turbo-0125:9HlXk9TZ (1 época de *fine-tuning*)	71.38 %
gpt-3.5-turbo-0125:9KV79yHU (2 épocas de *fine-tuning*)	53.33 %

NOTA: No se encontró información referente a la precisión de DISC-LawLLM

El modelo gpt-3.5-turbo-0125:9HlXk9TZ, ajustado durante una sola época, predijo correctamente el 71.38 % de todas las predicciones realizadas sobre las clases objetivo. Este resultado sugiere que un *fine-tuning* breve puede conducir a una precisión bastante alta en la tarea de interés.

En contraste, el modelo gpt-3.5-turbo-0125:9KV79yHU, ajustado durante dos épocas, muestra una precisión ligeramente inferior del 53.33 %, esto puede deberse a una posible sobreajuste o a una falta de generalización del modelo después de un *fine-tuning* prolongado. Este hallazgo sugiere una posible tendencia al descenso, aunque no proporciona una certeza sobre el tipo específico de comportamiento, dado que los datos pueden evolucionar de diversas maneras, como lineal, exponencial, logarítmica, polinómica, entre otras.

Estos resultados destacan la importancia de considerar la precisión al evaluar el rendimiento de los modelos ajustados (*fine-tuned*) para determinar su capacidad para predecir verdaderos positivos. Además, expone la necesidad de asignar más fondos para investigar de manera más detallada si un aumento en las épocas de *fine-tuning* tiene un impacto significativo en la precisión del modelo.

6.4. Recuperación (*Recall*)

Cuadro 6: Recuperación por modelo

Modelo	Recuperación (*Recall*)
gpt-3.5-turbo-0125 (base)	1.12 %
gpt-3.5-turbo-0125:9HlXk9TZ (1 época de *fine-tuning*)	69.44 %
gpt-3.5-turbo-0125:9KV79yHU (2 épocas de *fine-tuning*)	58.33 %

NOTA: No se encontró información referente a la recuperación de DISC-LawLLM

En el contexto de los modelos ajustados (*fine-tuned*), los resultados de la recuperación son especialmente relevantes, ya que una alta tasa de recuperación asegura que el modelo pueda identificar la mayoría, si no todas, las instancias relevantes de la clase de interés.

El modelo base, por su parte, mostró una tasa de recuperación del 1.12 %, lo que indica una capacidad muy limitada para identificar las instancias relevantes en el conjunto de datos.

En contraste, el modelo gpt-3.5-turbo-0125:9HlXk9TZ, ajustado durante una sola época, exhibe una recuperación del 69.44 %. Este resultado sugiere que el modelo puede identificar casi el 70 % de todas las instancias relevantes de la clase de interés en el conjunto de datos. Esta alta tasa de recuperación es alentadora, ya que indica que el modelo tiene una capacidad considerable para detectar casos relevantes, lo que es crucial en aplicaciones donde se requiere un alto grado de exhaustividad.

Por otro lado, el modelo gpt-3.5-turbo-0125:9KV79yHU, ajustado durante dos épocas, muestra una tasa de recuperación ligeramente inferior del 58.33 %. Aunque sigue siendo relativamente alta, puede interpretarse como una tendencia al descenso, sin embargo, no proporciona una certeza sobre el tipo específico de comportamiento, dado que los datos pueden evolucionar de diversas maneras, como lineal, exponencial, logarítmica, polinómica, entre otras. Por otra parte, esta cifra sugiere que el modelo puede estar perdiendo algunas instancias relevantes de la clase de interés en comparación con el modelo ajustado con una sola época. Esto puede deberse a una posible sobreajuste o a una falta de generalización del modelo después de un *fine-tuning* prolongado.

En conclusión, estos resultados subrayan la importancia de lograr una alta tasa de recuperación en los modelos ajustados (*fine-tuned*), ya que garantiza que el modelo pueda identificar de manera efectiva la mayoría de las instancias relevantes de la clase de interés. Una alta recuperación es fundamental en aplicaciones donde la exhaustividad es crucial, como la clasificación de textos legales o la detección de información importante en grandes volúmenes de datos. Finalmente, muestran la necesidad de asignar más fondos para investigar de manera más detallada si un aumento en las épocas de *fine-tuning* tiene un impacto significativo en la recuperación del modelo.

6.5. Validación del sistema gestión de consultas

6.5.1. Creación y envío de *prompts*

Para validar la ejecución correcta y esperada del algoritmo que forma parte del flujo del sistema de gestión, se llevaron a cabo los siguientes pasos:

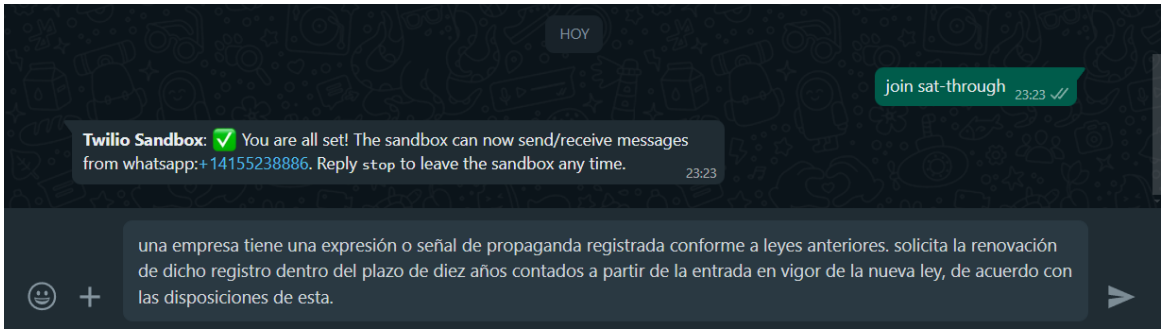


Figura 5: Conexión establecida y mensaje preparado para su envío

Inicialmente, identificamos el número telefónico de pruebas al que sería necesario conectarse. Después de registrarlo en WhatsApp, se ejecutó el comando de activación para que el *listener* del proveedor (Twilio) reconociera la solicitud de conexión. Una vez ingresado, se recibió un mensaje de confirmación de la conexión, que sirvió como validación de la primera etapa relacionada a la conectividad.

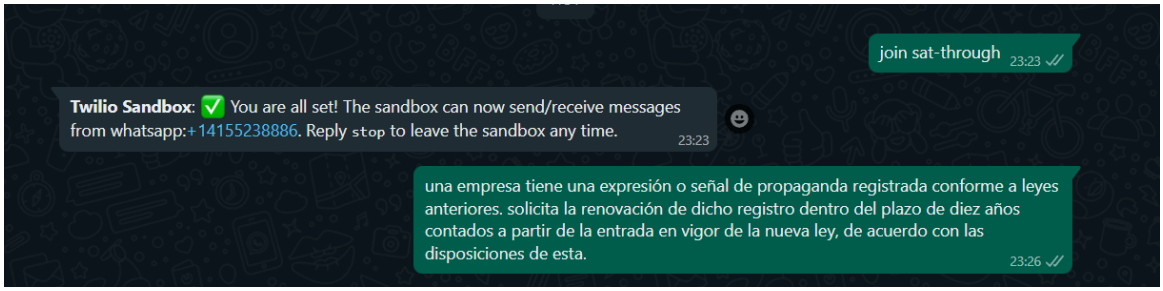


Figura 6: Mensaje enviado al número de prueba

Una vez establecida la conexión se procedió al envío directo del caso (dado que estamos corroborando funcionamiento del sistema, no de los algoritmos de asistencia), quedando a la espera de la respuesta.

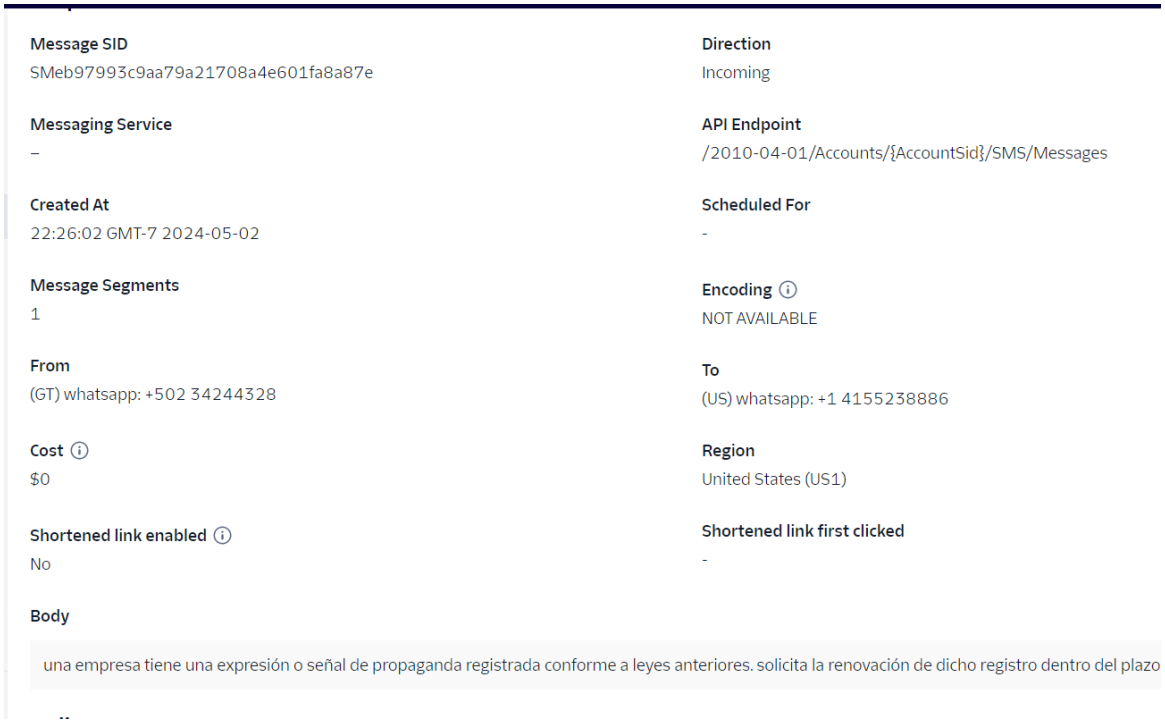


Figura 7: Recepción del mensaje en Twilio

Mientras se espera la recepción del mensaje, se verificó la llegada exitosa del mensaje al aplicativo proveedor de servicios SMS (Twilio), validando la segunda parte del flujo correspondiente a la recepción por parte de Twilio.

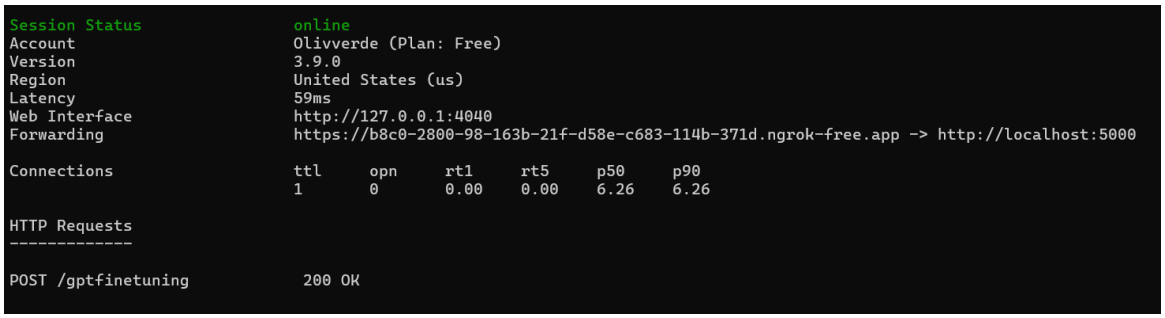


Figura 8: Recepción ingresante del mensaje en el tunel seguro

Una vez interceptado el mensaje entrante en Twilio, este debe de redireccionarlo al *microframework web* Flask, que funciona como servidor intermedio. Sin embargo, para llegar a este punto es necesario que se traslade por medio de un túnel seguro. En la figura 8 se puede observar que el túnel recibió el mensaje por parte de Twilio, retornando un código de ejecución exitosa (**200 OK**). Validando el correcto flujo del sistema hasta la tunelización.

```
ANAS/runner.py
 * Serving Flask app 'runner'
 * Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on all addresses (0.0.0.0)
 * Running on http://127.0.0.1:5000
 * Running on http://192.168.1.197:5000
Press CTRL+C to quit
una empresa tiene una expresión o señal de propaganda registrada conforme a leyes anteriores. solicita la renovación de dicho registro dentro del plazo de diez años contados
 a partir de la entrada en vigor de la nueva ley, de acuerdo con las disposiciones de esta.
```

Figura 9: Ingreso del mensaje al *microframework web* (servidor local) Flask

Ya dentro de la tunelización, el mensaje llegaría seguro al servidor intermedio (Flask). Si el mensaje llegó sin inconvenientes, se mostrará en la terminal el mensaje. El algoritmo contactaría el API de OpenAI para conectar con el modelo ajustado (*fine-tuned*) utilizando el formato JSON con los roles de sistema y usuario.

Con la impresión del mensaje en consola, se validó la ejecución del flujo hasta el servidor intermedio.

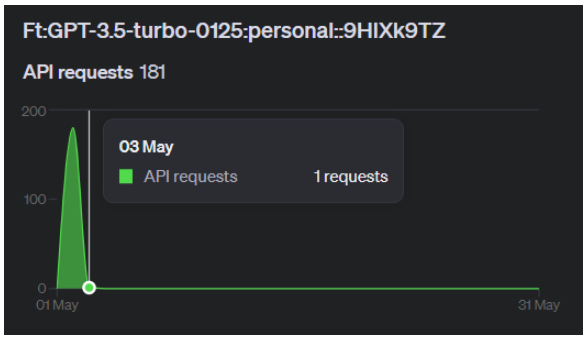


Figura 10: Ingreso API OpenAI e interacción con el modelo ajustado (*fine-tuned*)

Una vez enviado el *request* al API, el API lo intercepta listo para envíar la respuesta del modelo ajustado (*fine-tuned*). Una vez recibido por OpenAI se validó a través de su interfaz de *API requests* para el modelo gpt-3.5-turbo-0125:9HlXk9TZ.

```
 * Running on all addresses (0.0.0.0)
 * Running on http://127.0.0.1:5000
 * Running on http://192.168.1.197:5000
Press CTRL+C to quit
una empresa tiene una expresión o señal de propaganda registrada conforme a leyes anteriores. solicita la renovación de dicho registro dentro del plazo de diez años contados
 a partir de la entrada en vigor de la nueva ley, de acuerdo con las disposiciones de esta.
titulo 7 capitulo 1 y 2 no seccion : disposiciones transitorias y finales para la implementacion de la ley de propiedad industrial en guatemala.

Creo que el fragmento de ley encontrado en el Título 7, Capítulo 1 y 2, que trata sobre disposiciones transitorias y finales para la implementación de la ley de propiedad in
dustrial en Guatemala, se relaciona con el caso presentado porque establece las normas y procedimientos especiales que deben seguir aquellas empresas que tenían registros co
nforme a leyes anteriores y buscan renovarlos bajo la nueva ley. En este caso, la empresa que solicita la renovación de su registro de expresión o señal de propaganda dentro
 del plazo de diez años desde la entrada en vigor de la nueva ley estaría amparada por estas disposiciones transitorias y finales para llevar a cabo el proceso de renovación
 bajo las nuevas regulaciones establecidas en la ley de propiedad industrial.
127.0.0.1 - - [02/May/2024 23:26:08] "POST /gptfinetuning HTTP/1.1" 200 -
```

Figura 11: Respuesta del modelo ingresando al *microframework web* (servidor local) Flask

El mensaje viaja de vuelta al servidor intermedio (Flask), interceptando la respuesta en la terminal. Validando así la ejecución del flujo de respuesta desde OpenAI.

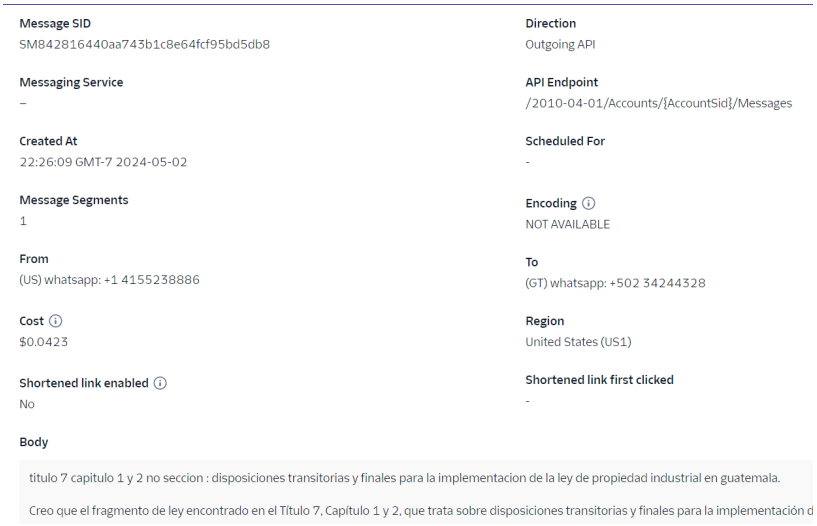


Figura 12: Respuesta del modelo ajustado (*fine-tuned*) recibida por Twilio por medio de túnel seguro

Posteriormente, el mensaje viaja de vuelta al proveedor de mensajes por medio de la tunelización. Se valida la llegada y su envío inmediato al chat de WhatsApp a través de los registro *Outgoing*

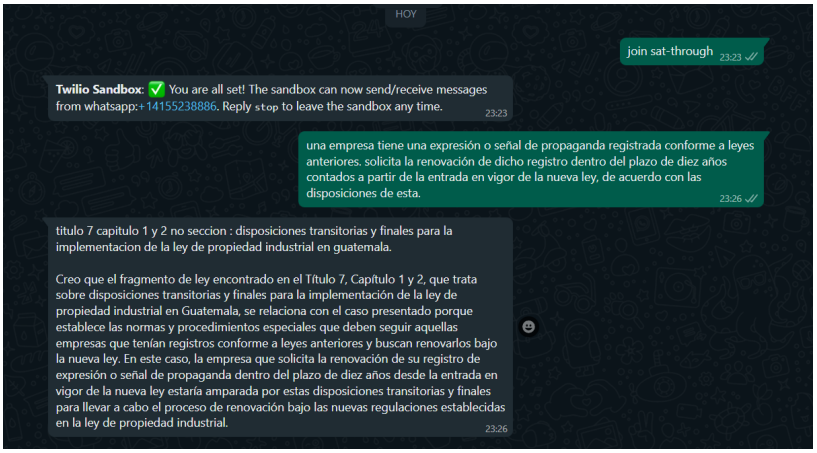


Figura 13: Respuesta del modelo ajustado (*fine-tuned*) al usuario por WhatsApp

Finalmente, el mensaje retorna al usuario emisor con la respuesta. Validando por completo la ejecución exitosa del flujo propuesto para el sistema.

El análisis del flujo de creación y envío de *prompts* reveló un desarrollo sólido en todas las fases de comunicación del sistema. Desde la interacción inicial del usuario con el sistema a través de WhatsApp hasta el procesamiento de la consulta por parte del modelo ajustado (*fine-tuned*), se observó una coherencia notable en la transmisión de mensajes a través de la infraestructura. Este flujo sin interrupciones demostró una implementación efectiva de la arquitectura propuesta, lo que contribuye a la confianza en la capacidad del sistema para facilitar una experiencia fluida y eficiente para los usuarios.

6.5.2. Guiamiento del usuario en la creación de *prompt* e interpretación de respuestas

Para validar el correcto funcionamiento del sistema referente a la orientación del usuario en la

creación de *prompts* e interpretación de las respuestas, es fundamental evaluar la capacidad que tiene para redirigir la conversación hacia estos puntos de interés. Por lo que fue necesario probarlo en tres casos particulares:

Caso 1: Se ingresa un *prompt* que simula una conversación convencional

Si el sistema no identifica directamente una consulta, un ingreso malicioso o señales de solicitud de asistencia o ayuda, responderá con un saludo e inmediatamente establecerá que solo podrá ayudarte en temas de propiedad intelectual. Posteriormente, te ofrecerá la opción de solicitar ayuda. Una vez solicitada, desplegará información descriptiva del sistema, así como información pertinente para crear un *prompt* adecuado, un ejemplo de caso y la estructura generalizada utilizada para la formulación de los casos de entrenamiento. Finalmente, indicará el significado de la respuesta una vez que el usuario haya hecho un ingreso.

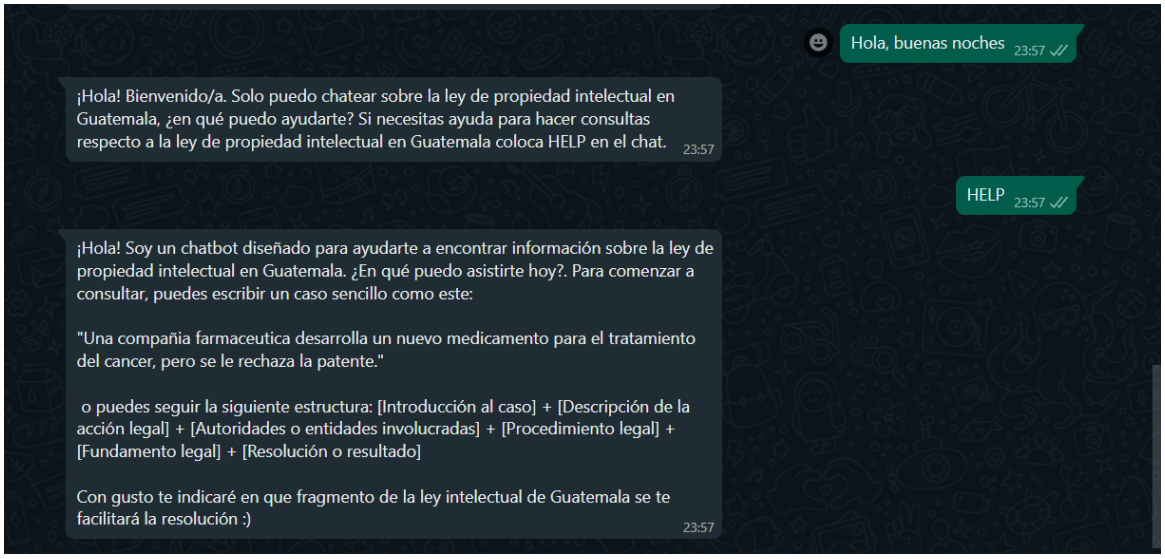


Figura 14: Mensaje de ayuda que indica la construcción del *prompt* y cómo interpretar la respuesta

Caso 2: Se ingresa un *prompt* malicioso o fuera del contexto legal intelectual

Si el sistema identifica un ingreso malicioso o ajeno a la ley de propiedad intelectual, activará el algoritmo de negación de comunicación, indicando que no podrá hablar de otros temas distintos a la ley de propiedad intelectual, posteriormente intentará redirigir la conversación a la solicitud de asistencia mediante el ofrecimiento de ayuda para interactuar con el modelo.

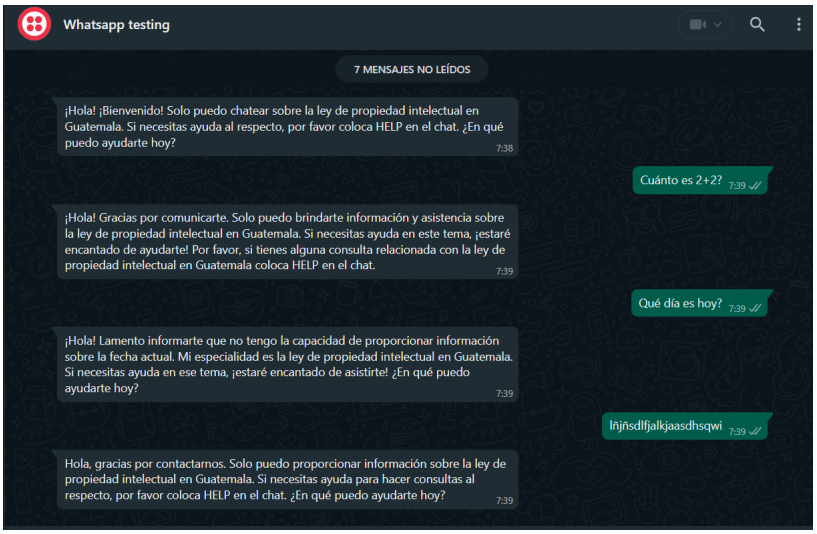


Figura 15: Algoritmo de negación de comunicación para casos externos a la ley de propiedad intelectual

Caso 3: Se ingresa un *prompt* para solicitar asistencia o ayuda

Si el sistema identifica un ingreso de solicitud de asistencia, efectuará el algoritmo de detección de asistencia de manera inmediata, desplegando información descriptiva del sistema, así como información pertinente para crear un *prompt* adecuado, un ejemplo de caso y la estructura generalizada utilizada para la formulación de los casos de entrenamiento. Finalmente, indicará el significado de la respuesta una vez que el usuario haya hecho un ingreso.

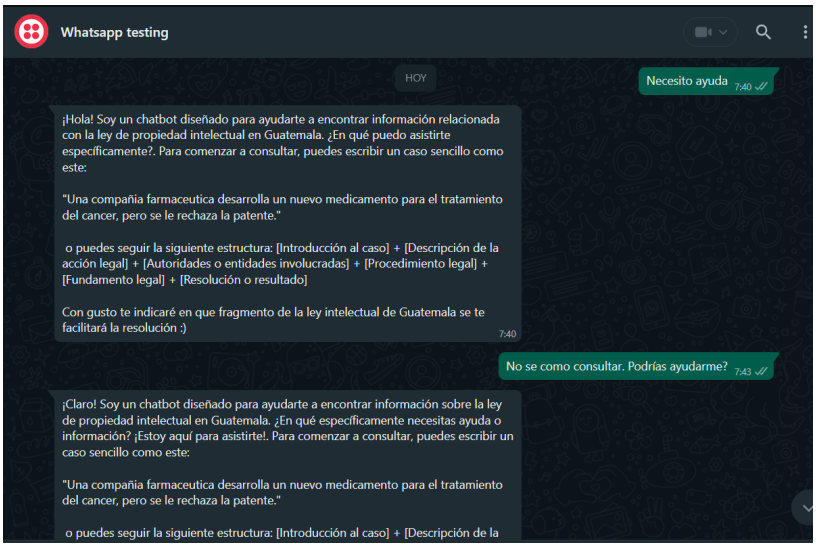


Figura 16: Algoritmo de detección de asistencia

La evaluación del guiamiento en la creación e interpretación de respuestas reveló un funcionamiento robusto de los algoritmos diseñados para asistir al usuario en ambos aspectos. Durante la interacción directa con el sistema, se observó cómo el algoritmo de detección de indicios de ayuda proporcionaba orientación efectiva sobre cómo estructurar un *prompt* adecuadamente. Asimismo, se evidenció que el sistema ofrecía información clara y concisa sobre el significado de la salida del

modelo ajustado (*fine-tuned*), lo que facilitaba la interpretación de las respuestas por parte del usuario. Esta funcionalidad demostró una implementación coherente de los algoritmos de guiamiento, contribuyendo así a mejorar la experiencia del usuario y a garantizar una interacción efectiva con el sistema de gestión de consultas.

7. Conclusiones

- Se logró realizar un LLM *fine-tuning* usando el modelo gpt-3.5-turbo-0125 con 1 y 2 épocas, demostrando capacidad para entender y responder consultas legales respecto a la ley de propiedad intelectual de Guatemala.

- El modelo ajustado (*fine-tuned*) de una época logró superar la exactitud del modelo base en aproximadamente 14 veces, la recuperación en 62 veces y la precisión en 6 veces, además de superar en un 70 % la exactitud del proyecto DISC-LawLLM basado en *fine-tuning* legal de patentes comerciales. Demostrando que el modelo ajustado (*fine-tuned*) de una época está equipado para comprender y responder consultas específicas sobre la ley de propiedad intelectual.

- El *fine-tuning* de una sola época superó al de dos épocas en todos los aspectos evaluados, incluyendo precisión (en un 33 %), exactitud (en un 19 %), recuperación (en un 34 %) y pérdida de validación completa (en un 8 %). Demostrando que una duración más larga del *fine-tuning* no asegura una mejora en el rendimiento.

- Se logró implementar un sistema de gestión de consultas con la capacidad de crear y enviar *prompts* desde WhatsApp, y obtener respuestas del modelo ajustado (*fine-tuned*).

- Se logró implementar una serie de algoritmos dentro del sistema de gestión de consultas, que permiten asistir al usuario en la formulación de *prompts* y la interpretación de las respuestas obtenidas por el modelo ajustado (*fine-tuned*).

- Se logró la implementación de una solución tecnológica basada en LLM *fine-tuning*, capaz de orientar al usuario en consultas sobre la ley de propiedad intelectual a través de la sinergia entre la eficiencia demostrada por el sistema de gestión en facilitar el envío, recepción y asistencia al usuario en la formulación de consultas, así como en la interpretación de las respuestas, y la habilidad del modelo ajustado (*fine-tuned*) para comprender y responder adecuadamente a las consultas legales sobre la ley de propiedad industrial de Guatemala.

8. Recomendaciones

- Se recomienda aumentar la cantidad de datos de entrenamiento y prueba para mejorar la diversidad de contextos y casos, lo que contribuirá a mejorar el rendimiento general del modelo.

- En caso de considerar el aumento del número de épocas de entrenamiento, se sugiere contar con un mayor número de observaciones o explorar la utilización de un modelo base diferente para obtener resultados más concluyentes.

- Si se planea implementar un sistema o interfaz que exponga el modelo ajustado (*fine-tuned*), se aconseja realizar pruebas de usabilidad y evaluación de la experiencia del usuario para determinar la aceptación y percepción del producto final.

- Para continuar con el enfoque en las leyes de propiedad intelectual en Guatemala, se recomienda considerar una segmentación más específica, lo que podría resultar en respuestas más precisas. Asimismo, se sugiere explorar enfoques adyacentes, como el reglamento de la ley de propiedad industrial, la clasificación de Niza, la ley de derechos de autor y derechos conexos, y los procedimientos de patentes, con el fin de ofrecer una visión más completa del panorama legal.

- Se aconseja experimentar con otros modelos dentro de la gama GPT y explorar otras opciones disponibles en el mercado para ampliar las posibilidades y mejorar el rendimiento del sistema.

- Con el propósito de facilitar la tarea de clasificación, se recomienda establecer contacto con profesionales del ámbito legal vinculados a instituciones educativas, con el fin de utilizar bases de datos previamente disponibles. Además, se sugiere realizar un análisis comparativo (*benchmarking*) con otros proyectos similares centrados en la resolución de exámenes legales oficiales, dado su enfoque y conveniencia legal comparables al presente proyecto.

9. Bibliografía

Referencias

Amazon. (2023). *¿Qué son los transformadores en la inteligencia artificial?* Amazon. https://aws.amazon.com/es/what-is/transformers-in-artificial-intelligence.

Dhruv P. (2022, 18 de mayo). *Understanding Micro Frameworks — FastAPI Flask; Pros, Cons, and Comparison* LinkedIn. https://www.linkedin.com/pulse/understanding-micro-frameworks-fastapi-flask-pros-cons-dhruv-patel/.

GeeksforGeeks. (2023). *Explanation of BERT Model – NLP* GeeksforGeeks. https://www.geeksforgeeks.org/explanation-of-bert-model-nlp/.

Google. (2022a, 26 de septiembre). *Clasificación: Exactitud* Google. https://developers.google.com/machine-learning/crash-course/classification/accuracy?hl=es-419.

Google. (2022b, 26 de septiembre). *Clasificación: Exactitud* Google. https://developers.google.com/machine-learning/crash-course/classification/accuracy?hl=es-419.

IBM. (2022, 1 de diciembre). *Utilización del túnel seguro para acceder a los recursos en una red privada* IBM. https://www.ibm.com/docs/es/cloud-paks/cp-management/2.3.x?topic=sa-using-secure-tunnel-access-resources-in-private-network.

Khawaja, R. (2023, septiembre). *Fine Tuning LLMs 101.* LinkedIn. https://datasciencedojo.com/blog/fine-tuning-llms/.

Koshti H. (2023, 28 de marzo). *Understanding the GPT-3.5 Architecture!.* LinkedIn. https://www.linkedin.com/pulse/chatgpts-guide-understanding-gpt-35-architecture-heena-koshti/.

Lambert, N. (2022, 9 de diciembre). *Illustrating Reinforcement Learning from Human Feedback (RLHF)* Hugging Face. https://huggingface.co/blog/rlhf.

Lark Editorial Team. (2023a, 27 de diciembre). *Parameters* Lark. https://www.larksuite.com/en_us/topics/ai-glossary/parameters.

Lark Editorial Team. (2023b, 26 de diciembre). *Supervised Fine Tuning* Lark. https://www.larksuite.com/en_us/topics/ai-glossary/supervised-fine-tuning.

Meta. (2023). *Introducing LLaMA: A foundational, 65-billion-parameter large language model* Meta. https://ai.meta.com/blog/large-language-model-llama-meta-ai/.

Multimodal. (2023, 13 de septiembre). *LLM Fine-Tuning: How To Choose the Right Model?* Multimodal. https://www.multimodal.dev/post/llm-fine-tuning.

NVIDIA. (2023, mayo). *What are Large Language Models?.* NVIDIA. https://www.nvidia.com/en-us/glossary/large-language-models/.

OpenAI. (2023a). *Fine-tuning* OpenAI. https://platform.openai.com/docs/guides/fine-tuning.

OpenAI. (2023b, diciembre). *GPT-3.5 Turbo* OpenAI. https://platform.openai.com/docs/models/gpt-3-5-turbo.

Pinecone. (2022). *Unsupervised Training for Sentence Transformers* Pinecone. https://www.pinecone.io/learn/series/nlp/unsupervised-training-sentence-transformers/.

Registro de la Propiedad Intelectual de Guatemala. (2000). *Guía General de Uso* Registro de la Propiedad Intelectual de Guatemala. https://portal.rpi.gob.gt/wp-content/uploads/2020/12/guiadelusuario.pdf.

Saxena S. (2018, 11 de mayo). *Precision vs Recall* Medium. https://medium.com/@shrutisaxena0617/precision-vs-recall-386cf9f89488.

Simplilearn. (2023, 7 de noviembre). *What is Epoch in Machine Learning?* Simplilearn. https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-epoch-in-machine-learning.

Sirit A. (2023, 9 de abril). *Que es el Prompting y como se usa para la Inteligencia Artificial* LinkedIn. https://www.linkedin.com/pulse/que-es-el-prompting-y-como-se-usa-para-la-artificial-alejandro-sirit/.

SuperAnnotate. (2024, 5 de febrero). *Fine-tuning large language models (LLMs) in 2024* SuperAnnotate. https://www.superannotate.com/blog/llm-fine-tuning.

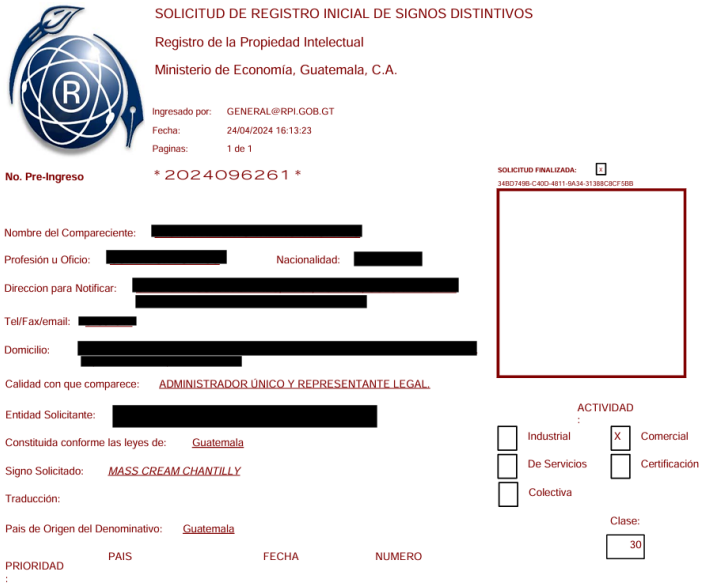
Toolify.ai. (2024, 8 de enero). *Discover the Cost of Fine-Tuning Flan-T5 LLM Models* Toolify.ai. https://www.toolify.ai/ai-news/discover-the-cost-of-finetuning-flant5-llm-models-411171.

Turing. (2023). *Fine-Tuning LLMs : Overview, Methods, and Best Practices* Turing. https://www.turing.com/resources/finetuning-large-language-models.

World Intellectual Property Organization. (2020). *What is Intellectual Property?* World Intellectual Property Organization. https://www.wipo.int/about-ip.

10. Anexos

10.1. Ejemplo de casos usados para creación de estructura generalizadas



SOLICITUD DE REGISTRO INICIAL DE SIGNOS DISTINTIVOS

Registro de la Propiedad Intelectual

Ministerio de Economía, Guatemala, C.A.

Ingresado por: GENERAL@RPI.GOB.GT
Fecha: 24/04/2024 16:13:23
Paginas: 1 de 1

No. Pre-Ingreso * 2024096261 *

SOLICITUD FINALIZADA: X
34BD749B-C40D-4811-9A34-31388CBCF5BB

Nombre del Compareciente: [redacted]
Profesión u Oficio: [redacted] Nacionalidad: [redacted]
Direccion para Notificar: [redacted]
Tel/Fax/email: [redacted]
Domicilio: [redacted]
Calidad con que comparece: ADMINISTRADOR ÚNICO Y REPRESENTANTE LEGAL.
Entidad Solicitante: [redacted]
Constituida conforme las leyes de: Guatemala
Signo Solicitado: *MASS CREAM CHANTILLY*
Traducción:
Pais de Origen del Denominativo: Guatemala

PRIORIDAD
:
PAIS FECHA NUMERO

ACTIVIDAD
:
[ ] Industrial [X] Comercial
[ ] De Servicios [ ] Certificación
[ ] Colectiva

Clase: 30

Figura 17: Caso 1 solicitud de patente comercial Mass Cream (1/2)

1) Concretar mercancías, actividades o servicios que ampara:

PRODUCTOS DE PASTELERIA, CONFITERIA Y HELADOS.

2) Reservas y/o Renuncias:

EL TITULAR SE RESERVA EL USO SIN RESTRICCIONES NI LIMITACIONES A CUALQUIER TAMAÑO, EN LA COMBINACION DE DISTINTOS COLORES, PARA APLICARLO SOBRE PRODUCTOS DE PASTELERIA, CONFITERIA Y HELADOS, ASI COMO BOLSAS, CONTENDORES, ACCESORIOS Y PAPELERIA EN GENERAL. SE RESERVA LA DIFUSION A CUALQUIER MEDIO PUBLICITARIO, VISUAL, AUDITIVO, MEDIANTE RADIO, TELEVISION, INTERNET Y OTROS MEDIOS SIMILARES.

3) Dirección del lugar principal en que se fabriquen, distribuyan, comercialicen o presten los productos o servicios

[redacted]

Acompaño a la solicitud:

[X] 4 Reproducciones [X] Nombramiento [ ] Poder
[ ] Fotocopia de DPI [X] Comprobante de pago de Tasa

PASAPORTE Y PATENTES.

Lugar y Fecha: GUATEMALA, 24 de abril de 2024

(f):_____
Compareciente

En su auxilio:_____
Firma y sello del Abogado

Figura 18: Caso 1 solicitud de patente comercial Mass Cream (2/2)

El caso en cuestión destaca como un ejemplo representativo en el campo de la propiedad intelectual. Mass Cream, una empresa dedicada a la producción de productos pasteleros, específicamente crema chantilly, se encuentra en proceso de registro de su marca. Este procedimiento implica cumplir con requisitos administrativos y someter la marca a evaluación conforme a las regulaciones pertinentes, que abarcan aspectos visuales, terminológicos y promocionales, tales como imágenes, palabras y eslóganes.

Actualmente, este proceso sigue en curso, permitiendo la generación de algunos escenarios genéricos como:

- **CASO:** Un individuo está interesado en registrar una marca para su nuevo producto, pero está confundido sobre qué tipo de signos pueden constituir una marca según la Ley de Propiedad Intelectual. ¿Pueden las palabras solas ser consideradas marcas?
    - **ETIQUETA: título 2 capítulo 1 sección 1: marcas: normas, adquisicion, prioridad, derechos, inadmisibilidad en guatemala**
- **CASO:** Una marca denominativa tiene un elemento con significado en otro idioma. ¿Se necesita una traducción simple?
    - **ETIQUETA: título 2 capítulo 1 sección 2: registro de marcas: requisitos, procedimientos y efectos legales específicos.**

10.2. GitHub proyecto

Figura 19: Pantalla principal GitHub proyecto LLM-FINE-TUNNING-PI

Link de enlace: https://github.com/Olivverde/LLM-FINE-TUNING-PI.git