

**Sistema De QA Basado en los Informes del Conflicto Armado En Colombia**

Trabajo De Grado

JUAN DAVID TOVAR MONTOYA

202029032

[juan.david.tovar@correounivalle.edu.co](mailto:juan.david.tovar@correounivalle.edu.co)

Nombre del Director

RAUL E. GUTIERREZ DE PIÑEREZ REYES, Ph.D

[raul.gutierrez@correounivalle.edu.co](mailto:raul.gutierrez@correounivalle.edu.co)

Facultad de Ingeniería

Escuela de Ingeniería de Sistemas y Computación

Programa Académico de Ingeniería de Sistemas

Cali, Mayo 30 2024

**Índice**

[**Resumen 2**](#_i88ckgltluo1)

[**Planteamiento del Problema 3**](#_6nvgm11t0xdr)

[**Justificación 5**](#_v1koz0ehhiet)

[**Objetivos 6**](#_xr15mmon74cb)

[Objetivo General 6](#_xo213zmh0pdk)

[Objetivos Específicos 6](#_807fr349nok4)

[Resultados Esperados 7](#_nvn4g4eqhukl)

[**Alcance 8**](#_7psgbdtzqy)

[Límites del proyecto: 8](#_10fdd2qd436i)

[Aspectos no abordados: 9](#_81jacvhe80gc)

[Consideraciones adicionales: 9](#_p9ut4765f2uk)

[**Marco Referencial 11**](#_5aq2suk448ex)

[Glosario 11](#_u11ql0ittbim)

[Marco Teórico 12](#_tj1wrs3w5wkq)

[Conflicto Armado en Colombia 12](#_b0x65x5ndojk)

[Tensores, Transformers y El Modelo de Atención 14](#_ne0jojmd6ft)

[Procesamiento del Lenguaje Natural (PLN) y los LLM 15](#_gaj9o6xo4dx3)

[Los Sistemas de QA, Chatbots y el Fine-tuning de Modelos 16](#_idzjti4orpup)

[Datasets De Entrenamiento 18](#_agtd3lzgj3uu)

[Plataforma HuggingFace 19](#_6jqkgalmvntf)

[Antecedentes 20](#_4lop6mvvfv6a)

[Procesamiento del Lenguaje Natural Por Parte de La Comisión de La Verdad 20](#_s1392jwmp11w)

[Google Gemini - Chatbot con Datos de la web 21](#_rslbr4h8ue81)

[Modelo Entrenado Con Historias Del Conflicto 22](#_858w2dafa65g)

[Dataset de Pregunta-Respuesta en Español 23](#_fyaxhtpmfgyd)

[Conclusión Antecedentes 23](#_46rggd7ygqfv)

[**Metodología 25**](#_owq3jpxrdr0c)

[Actividades 25](#_si4wfkieod50)

[Actividad relacionada con cada objetivo. 25](#_jk85sqee504g)

[Cronograma de Actividades 27](#_8ber17jalcl9)

[Presupuesto 28](#_pfpxpoj2dgqs)

[**Bibliografía 29**](#_d3pnd4is18rs)

# 

# **Resumen**

Este trabajo de grado se enfoca en el desarrollo de un sistema de pregunta-respuesta (QA) que facilite el acceso y comprensión de la información contenida en los informes de la Comisión de la Verdad sobre el conflicto armado colombiano y el proceso de paz. El sistema utilizará tecnologías de inteligencia artificial como los grandes modelos de lenguaje (LLM) y el fine-tuning de modelos.

Se desarrollará un dataset anotado con preguntas y respuestas basadas en los informes de la Comisión de la Verdad. Luego, se implementarán técnicas de fine-tuning para ajustar un LLM de generación de texto a este dataset. El modelo afinado se integrará en la plataforma HuggingFace para ser probado y evaluado.

El objetivo principal es facilitar el acceso a la información sobre el conflicto armado colombiano y el proceso de paz de manera rápida y eficiente. Esto promoverá la investigación, el análisis y la comprensión de estos temas cruciales para la reconciliación en el país.

El sistema QA resultante será una herramienta valiosa para investigadores, estudiantes y público en general interesado en comprender la historia del conflicto armado colombiano y contribuir a la construcción de una paz duradera.

# **Planteamiento del Problema**

El conflicto armado colombiano, que ha durado más de cinco décadas, dejó un legado de dolor, sufrimiento y miles de víctimas. La Comisión de la Verdad, creada en el marco del Acuerdo de Paz firmado en 2016, tiene como objetivo esclarecer los hechos y contribuir a la construcción de una paz duradera en el país. Para cumplir con este mandato, la Comisión ha publicado una serie de informes que contienen información valiosa sobre el conflicto, sus actores y las víctimas. Como queda en el Decreto 588 de 2017, artículo 2º, uno de los objetivos de la comisión es contribuir al esclarecimiento de lo ocurrido [(Legislación Colombiana, n.d.)](#bmhhcdhqc6iq). Los informes son una herramienta fundamental para cumplir con este objetivo. Son documentos extensos y detallados que sintetizan los hallazgos de la Comisión sobre el conflicto armado. Estos documentos contienen información invaluable para comprender la historia del país y promover la reconciliación. Sin embargo, a pesar de su importancia, existe un desafío significativo: en Colombia, los índices de lectura de libros son bajos, alrededor del 64,0% lee mayoritariamente contenido de redes sociales (DANE, 2017) lo que limita el alcance y el impacto de estos informes en la población general. En contraste, se observa un uso creciente de tecnologías modernas como los chatbots para acceder a información, consultar información preguntando a Chatgpt puede ser hasta un 65.2% más eficiente que buscar en Google (Xu et al., 2023, p.12). Esta realidad plantea la necesidad de adaptar la difusión de estos importantes informes a formatos más accesibles y contemporáneos, como un sistema pregunta-respuesta utilizando interfaces de chat, para facilitar su comprensión y ampliar su alcance entre la ciudadanía.

Existen diversas metodologías para desarrollar sistemas de pregunta-respuesta (QA) con respuestas en lenguaje natural. Sin embargo, crear un sistema QA especializado en un tema específico presenta un desafío significativo. El procesamiento del lenguaje natural (PLN) es un campo de la inteligencia artificial que se dedica al estudio y desarrollo de sistemas que pueden analizar, comprender y generar lenguaje natural de forma automática. Estos sistemas han experimentado un progreso notable en los últimos años, impulsados por el auge de los grandes modelos de lenguaje (LLM) como GPT-3, BERT y Llama. La adopción de tecnologías de inteligencia artificial, como los LLM y el fine-tuning de modelos, presenta una oportunidad para abordar estos desafíos de manera innovadora. Estas tecnologías permiten desarrollar sistemas inteligentes que pueden procesar grandes volúmenes de texto, comprender su contenido y ofrecer respuestas relevantes y contextualmente apropiadas.

Por lo tanto, este trabajo de grado se enfoca en abordar la siguiente pregunta: ¿Cómo puede la aplicación de tecnologías de inteligencia artificial, como los LLM y el fine-tunning de modelos, aportar al desarrollo de un sistema QA que facilite el acceso, comprensión e interacción con la información contenida en los informes de la Comisión de la Verdad?

## 

## 

# **Justificación**

El sistema QA permitirá a los investigadores y estudiantes acceder a información sobre el conflicto armado colombiano y el proceso de paz de manera rápida y eficiente. Esto facilitará la investigación y el análisis de estos temas. El desarrollo del sistema QA permitirá la aplicación de diversos conceptos y técnicas de inteligencia artificial (IA), como procesamiento del lenguaje natural (PLN) y aprendizaje automático (ML) . El proyecto puede contribuir al avance del conocimiento en estas áreas, ya sea mediante la implementación de nuevas técnicas o la mejora de las existentes. La documentación del proyecto puede servir como referencia para futuros investigadores que trabajan en el desarrollo de sistemas QA o chatbots.

La implementación de un sistema QA basado en inteligencia artificial puede reducir significativamente el tiempo y los recursos necesarios para acceder y analizar la información contenida en los informes de la Comisión de la Verdad. Esto se traduce en ahorros de costos tanto para investigadores individuales como para instituciones académicas y organizaciones no gubernamentales que trabajan en temas relacionados con el conflicto armado y la paz.

El resultado final de este proyecto tendrá un impacto social, ya que permitirá a la sociedad colombiana acceder a información sobre el conflicto armado y el proceso de paz, de una forma más rápida y concreta.

# **Objetivos**

## Objetivo General

Desarrollar un sistema de pregunta-respuesta (QA) que facilite el acceso y comprensión de la información contenida en los informes de la Comisión de la Verdad sobre el conflicto armado colombiano y el proceso de paz, utilizando tecnologías de inteligencia artificial como los grandes modelos de lenguaje (LLM) y el fine-tuning de modelos.

## **Objetivos Específicos**

1. Desarrollar un dataset anotado para QA con base en la información de los informes de la comisión de la verdad. Teniendo en cuenta que son alrededor de 700 informes de los que se pueden sacar 3 preguntas básicas especificadas en la [temática del alcance](#gpir24lt5q3r) de este proyecto, la anotación del dataset se hará con la supervisión de un experto.
2. Implementar las técnicas de fine-tuning con el fin de ajustar un LLM de generación de texto usando un dataset anotado de QA sobre el tema de los informes de La Comisión De La Verdad.
3. Integrar el modelo afinado a la plataforma HuggingFace con el fin de testear el modelo con la interfaz y obtener el pipeline proporcionados por la plataforma. HuggingFace será una parte importante del sistema. HuggingFace proporciona interfaz gráfica y despliegue del modelo. Esto asegura una experiencia de usuario amigable, respuestas rápidas, y una gestión automática de seguridad y escalabilidad. HuggingFace maneja la infraestructura, permitiendo enfocarse en la mejora del modelo.
4. Implementar las medidas de desempeño del modelo afinado y análisis de resultados.

## Resultados Esperados

Tabla 1.

| OBJETIVO ESPECÍFICO | RESULTADO O PRODUCTO ESPERADO |
| --- | --- |
| 1 | Tener un dataset de pregunta-respuesta con al menos 2100 sentencias anotadas. Con preguntas relacionadas a temas específicos de los informes de la comisión. Serán 3 por cada informe.  (Número de Informes)x(Preguntas Básicas) = 2100 |
| 2 | Un modelo afinado en español que servirá de modelo pre-entrenado para sistemas de pregunta-respuesta en el tema. |
| 3 | Un modelo que se pueda probar mediante la interfaz de HuggingFace y un pipeline. |
| 4 | Resultados de perplejidad, precisión, fluidez y relevancia del modelo junto con métricas de tiempo de respuesta y uso de recursos. Además de un informe donde se analice las respuestas dadas por el modelo. |

# **Alcance**

## Límites del proyecto:

**Tipo de sistema QA:** Este proyecto se enfocará en el desarrollo de un sistema QA extractivo basado en texto. No se abordará el desarrollo de un sistema QA multimodal que incluya interacción con audio o video.

**Temática:** El sistema QA se centrará en la información contenida en los informes de la Comisión de la Verdad sobre el conflicto armado colombiano y el proceso de paz. Más de 15 informes finales (+500 páginas por informe) y cerca de 700 informes de casos.

Se consultarán fuentes complementarias confiables, como bases de datos oficiales, artículos académicos y reportes de organizaciones de derechos humanos, para ampliar y contextualizar la información. Los temas que se abordará son los siguientes:

1. Resumen de los hallazgos principales de cada informe.
2. Detalles sobre los casos y eventos específicos documentados en los informes.
3. Análisis de las causas estructurales del conflicto y las recomendaciones para la no repetición.

**Población objetivo**: El sistema QA estará dirigido a investigadores, estudiantes y público en general interesado en el conflicto armado colombiano y el proceso de paz. No se diseñará un sistema personalizado para un grupo específico de usuarios con necesidades particulares.

**Análisis de sesgos:** Se realizará un análisis de los posibles sesgos presentes en las respuestas generadas por el sistema de QA. Este análisis se llevará a cabo con la asesoría de un experto en el tema, garantizando que se identifiquen y mitiguen sesgos de manera adecuada. Este análisis se hará a medida que se entrena el modelo y también sobre el modelo final.

**Medidas Para Mitigar Sesgos:** Para desarrollar el modelo final que utilizará el sistema QA, se partirá de modelos pre-entrenados y actuales, tales como Llama 3 o Gemma. Estos modelos han sido previamente entrenados y ajustados con el fin de proporcionar respuestas neutrales. El proceso de ajuste (fine-tuning) se llevará a cabo con varios de estos modelos, y posteriormente, se realizará un análisis de sesgos. Este análisis permitirá identificar y seleccionar el modelo que presente el menor grado de sesgos. Además, durante la creación del dataset, se contará con la asesoría de un experto en la materia. Esta colaboración contribuirá significativamente a la mitigación de sesgos, asegurando que el modelo sea entrenado con respuestas más adecuadas y equilibradas.

## Aspectos no abordados:

**Desarrollo de un chatbot**: Si bien el proyecto contempla la creación de una interfaz de chat para el sistema QA, no se desarrollará un chatbot completo con funcionalidades avanzadas como la gestión de diálogos complejos o la personalización de la interacción con el usuario.

**Soporte para múltiples idiomas:** El sistema QA solo ofrecerá interacción y soporte en español.

**Medición del Impacto Social:** No se realizará una medición directa del impacto social del sistema QA. Sin embargo, se anticipa que el proyecto tendrá un impacto social indirecto al facilitar el acceso a información crucial sobre el conflicto armado y el proceso de paz en Colombia, lo cual podría contribuir a la educación y concienciación pública sobre estos temas.

## Consideraciones adicionales:

* El alcance del proyecto podría ajustarse en función de los recursos disponibles, el tiempo de desarrollo y las limitaciones técnicas.
* Se priorizará la calidad y el rigor metodológico en el desarrollo del sistema QA, a pesar de las limitaciones mencionadas.
* Se buscará la colaboración con expertos en diferentes áreas, como historia colombiana, conflicto armado y proceso de paz para enriquecer el proyecto. Tener la asesoría de un experto ayudará a crear un modelo con respuestas adecuadas y poco sesgadas.

Para concluir el alcance, este proyecto de grado busca desarrollar un sistema QA de calidad que facilite el acceso y la comprensión de la información contenida en los informes de la Comisión de la Verdad sobre el conflicto armado colombiano y el proceso de paz, utilizando tecnologías de inteligencia artificial de manera responsable y accesible.

# **Marco Referencial**

## Glosario

**IA**: Inteligencia Artificial

**LLM**: Large Language Models traducido al español como Grandes Modelos de Lenguaje.

**ML**: Machine Learning traducido al español como Aprendizaje Automático.

**PLN**: Procesamiento del Lenguaje Natural.

**QA**: Question Answering traducido al español como pregunta-respuesta.

**Lenguaje** **Natural**: La forma de comunicación que utilizan los humanos para interactuar entre sí. Se caracteriza por ser espontáneo, fluido y adaptable.

**Chatbot**: Un programa informático que simula una conversación con los usuarios [(La Web del Programador, n.d.)](#dc1zafsbsavk).

**Fine**-**tuning**: Es un enfoque para transferir el aprendizaje en el que un modelo previamente entrenado se entrena en datos nuevos [(Quinn et al., 2019, p. 551).](#w40lj9aq2b8n)

**Redes** **Neuronales** **Artificiales**: Son un campo fascinante de la inteligencia artificial que se inspira en el funcionamiento del cerebro humano [(Hardesty, 2017)](#j754p4xpzi3z).

**Dataset**: Colección de datos estructurados y enfocados para una tarea específica [(Universidad De Los Andes, 2024)](#zd5drg1fgdi).

## Marco Teórico

En este capítulo se presentan los conceptos básicos relacionados con el diseño y desarrollo de sistemas QA, enfocados en la consulta y análisis de grandes cantidades de datos expresados en lenguaje natural. Particularmente, se abordará la aplicación de estas tecnologías avanzadas en el contexto del conflicto armado en Colombia, basado en los informes de la Comisión de la Verdad.

### Conflicto Armado en Colombia

Con base al informe final "Hay futuro si hay verdad" (Comisión de la Verdad, 2022) se resumen varios hallazgos que dan contexto sobre el conflicto armado desde sus inicios. El conflicto armado en Colombia, según el informe, se extendió por más de cinco décadas, iniciando aproximadamente en 1964 y prolongandose hasta la firma del Acuerdo de Paz en 2016. Si bien el acuerdo marcó el fin del conflicto formal entre el Estado y las FARC-EP, la violencia y los enfrentamientos armados continúan en algunas regiones del país.

**¿Cuántas víctimas hubo?**

Determinar un número exacto de víctimas del conflicto armado en Colombia es complejo debido a la naturaleza prolongada y dispersa del mismo. Sin embargo, el informe estima que el conflicto ha dejado un saldo de más de 9 millones de víctimas, incluyendo:

* Más de 220.000 personas asesinadas.
* Más de 75.000 personas desaparecidas.
* Miles de víctimas de violencia sexual.
* Numerosas víctimas de secuestro, reclutamiento forzado y otros crímenes.

**¿Qué tipo de delitos se cometieron?**

El informe documenta una amplia gama de graves violaciones a los derechos humanos e infracciones al derecho internacional humanitario durante el conflicto armado, incluyendo:

* Asesinatos extrajudiciales, ejecuciones sumarias y masacres.
* Desapariciones forzadas.
* Desplazamiento forzado.
* Reclutamiento forzado de niños, niñas y adolescentes.
* Violencia sexual.
* Tortura.
* Atentados contra la población civil.
* Crímenes contra pueblos indígenas y afrodescendientes.
* ¿Quiénes fueron los más afectados?

El informe señala que los grupos más afectados por el conflicto armado fueron:

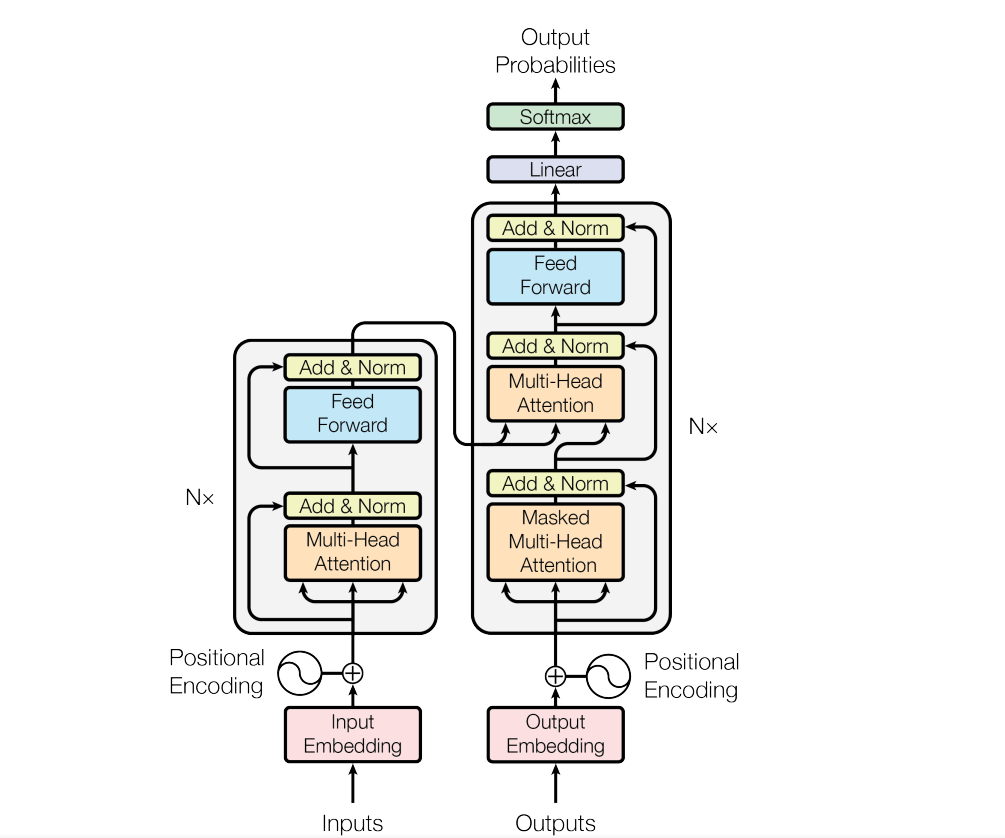
* Las comunidades rurales y campesinas, especialmente en zonas de alta presencia de grupos armados.
* Los pueblos indígenas y afrodescendientes.
* Las mujeres y las niñas, quienes han sido víctimas de violencia sexual y otras formas de violencia de género.
* Los niños, niñas y adolescentes, quienes han sido reclutados por grupos armados, han sufrido desplazamiento forzado y han sido víctimas de otras graves violaciones a sus derechos.
* Las personas LGBTI.
* Los defensores de derechos humanos y líderes sociales.

Es importante destacar que el informe también reconoce la participación de diversos actores en el conflicto, incluyendo al Estado, los grupos guerrilleros, los paramilitares y las organizaciones criminales. El informe enfatiza la necesidad de un proceso integral de verdad, justicia, reparación y no repetición para garantizar la construcción de una paz duradera en Colombia.

### Tensores, Transformers y El Modelo de Atención

Los tensores son estructuras matemáticas de gran relevancia, especialmente en el campo de la inteligencia artificial. Su capacidad para representar datos multidimensionales de manera eficiente los convierte en herramientas esenciales para el desarrollo y entrenamiento de redes neuronales. Estas redes neuronales suelen tener tareas como reconocimiento de imágenes, generación de contenido y toma de decisiones. Los tensores tienen una gran relevancia en los modelos de inteligencia artificial generativos de la actualidad. Es decir, modelos que generan contenido como GPT-3, Midjourney o Stable Diffusion. Por otro lado, los transformers son una arquitectura de red neuronal que ha revolucionado el campo del procesamiento del lenguaje natural (PLN) en los últimos años. Introducidos por Google Research en su artículo "Attention is All You Need", los transformers han demostrado un rendimiento superior en una amplia gama de tareas útiles en PLN. A diferencia de las redes neuronales tradicionales, que procesan el texto de manera secuencial, los transformers utilizan un mecanismo de atención que les permite "atender" a diferentes partes del texto de manera simultánea. Esto les permite capturar relaciones de largo alcance entre las palabras y frases, lo que resulta crucial para comprender el significado del texto [(Vaswani et al., 2017)](#er8xa1hjv3jl).

**Figura 1**



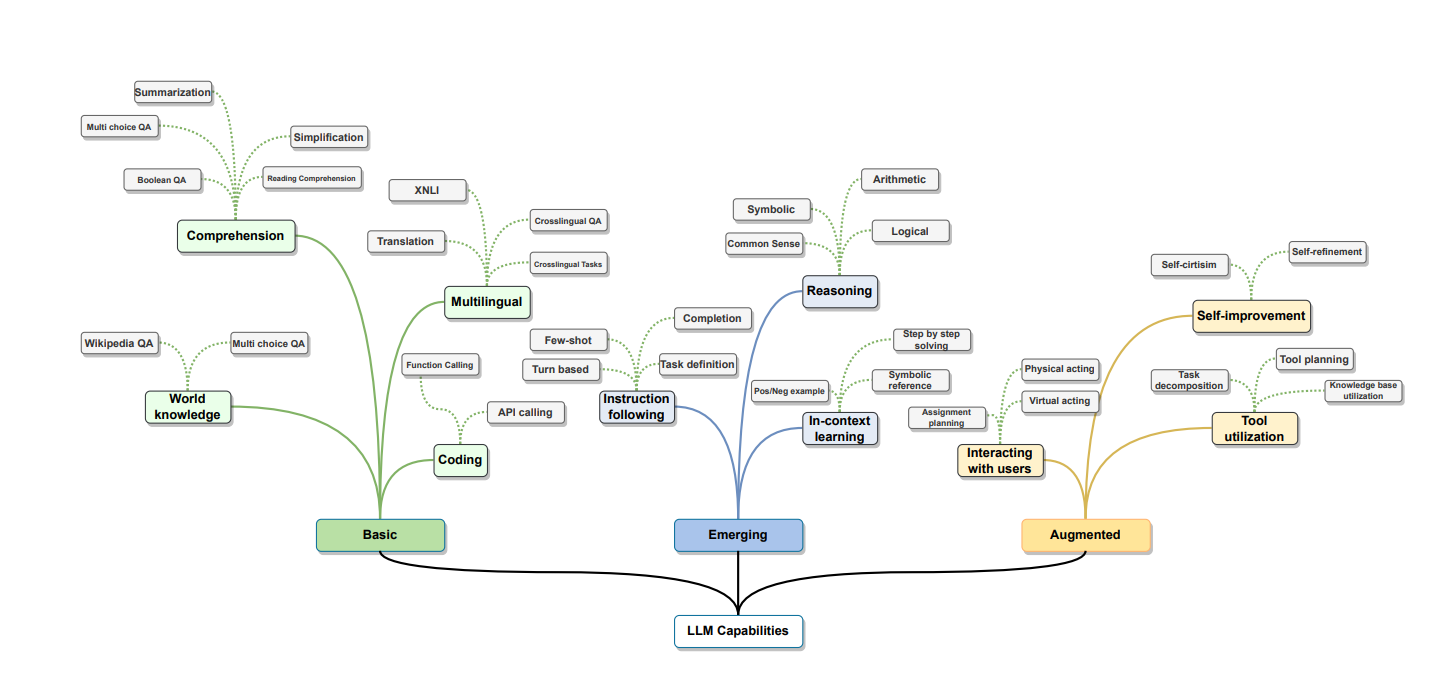
*Nota*. The Transformer - Model architecture [(Vaswani et al., 2017, p. 3)](#er8xa1hjv3jl) .

### Procesamiento del Lenguaje Natural (PLN) y los LLM

El procesamiento del lenguaje natural (PLN) es un campo de la inteligencia artificial que se encarga del estudio y desarrollo de sistemas capaces de analizar, comprender y generar lenguaje natural de manera automática [(Nadkarni, P. M & Et al, 2012)](#i3pngcd9qy22). Las tecnologías de PLN han experimentado un avance significativo en los últimos años, impulsadas por el auge de los LLMs como GPT-3, BERT y Llama.

Los LLMs son modelos de aprendizaje automático entrenados en grandes conjuntos de datos de texto. Estos modelos son capaces de aprender las relaciones estadísticas entre las palabras y frases, lo que les permite realizar tareas como la traducción automática, la generación de texto y la pregunta-respuesta (Minaee & Mikolov, 2024, p. 1). En la figura a continuación se pueden ver algunas de sus capacidades.

**Figura 2**



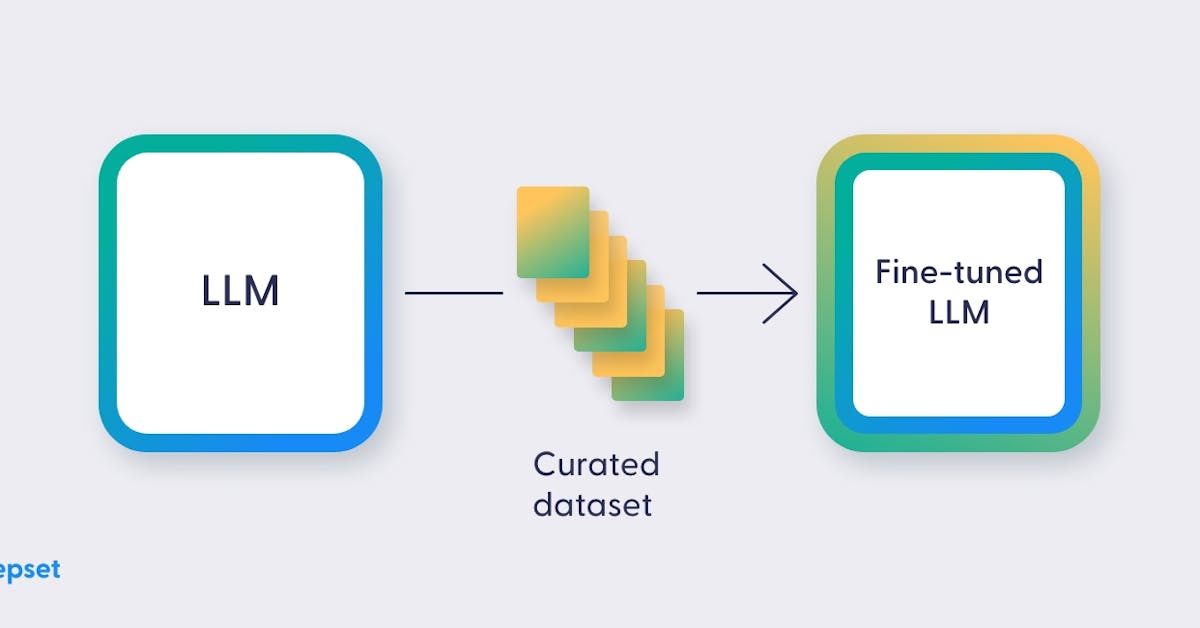
*Nota*. LLM Capabilities [(Minaee & Mikolov, 2024, p. 2)](#el6pnbva97b).

## 

### Los Sistemas de QA, Chatbots y el Fine-tuning de Modelos

El fine-tuning de modelos es una técnica de aprendizaje automático que consiste en ajustar un modelo pre-entrenado en una tarea específica [(Quinn et al. 2019, CAP 14)](#w40lj9aq2b8n). En el caso de los LLMs, el fine-tuning se puede utilizar para optimizar el modelo para responder preguntas sobre un tema específico, como el conflicto armado colombiano y el proceso de paz. Por otro lado, los chatbots son programas informáticos que simulan una conversación con los usuarios. Los chatbots inteligentes utilizan tecnologías de PLN con Inteligencia Artificial y buscan interactuar con los usuarios en distintas tareas siguiendo un flujo de conversación [(Kaur et al., 2024, p. 2)](#upvgg5n4lqfv).

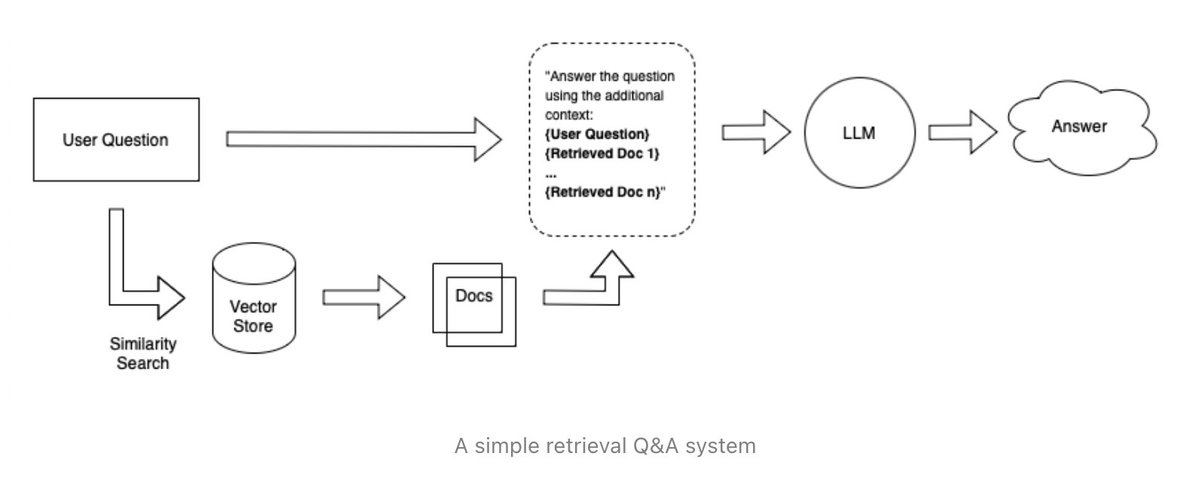
**Figura 3**



*Nota*. Resultado de Fine-tuning [(Deepset.ai, 2023)](#9qwdbfetnlm0).

En contraste con lo dicho, un sistema pregunta-respuesta, también conocido como Question Answering (QA), prioriza proporcionar respuestas precisas a preguntas específicas sin necesidad de un diálogo extenso. Los sistemas QA actuales utilizan grandes modelos de lenguaje (LLM) han revolucionado la forma en que interactuamos con la información. Estos sistemas, a diferencia de los sistemas QA tradicionales basados en reglas específicas para el texto de entrada [(Caballero, 2021, p.1 - p.2)](#xo5s4hwnjo22), aprovechan el poder de los LLMs para comprender y responder preguntas de manera más natural, precisa y completa.

**Figura 4**



*Nota*. Sistema QA con contexto extraído de documentos [(Chase, 2023).](#9ol54dsetgyb)

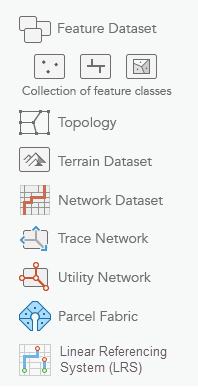
### Datasets De Entrenamiento

Un dataset es una colección organizada de datos, que pueden ser números, palabras, imágenes, sonidos, o cualquier otro tipo de información, que se utilizan como entrada para un modelo de aprendizaje automático [(H. Renear et al., 2011, p. 2)](#pb7q7wgq81x1). Un dataset típico consta de una serie de entradas (también llamadas muestras o ejemplos) que se componen de características o variables, y una etiqueta o respuesta asociada a cada entrada. En el caso de un sistema de pregunta-respuesta o Question Answering (QA), el dataset podría consistir en pares de preguntas y respuestas relacionadas con los informes de la Comisión de la Verdad. La calidad y cantidad de los datos en el dataset de entrenamiento serán fundamentales para el éxito del fine-tuning del LLM y la precisión del sistema de pregunta-respuesta [(Brazdil et al., 2022, p. 55)](#x287r7fd85y).

Otras características importantes a la hora de elaborar un dataset son:

* **Relevancia**: los datos deben ser relevantes para el problema que se va a resolver.
* **Calidad**: los datos deben ser precisos y consistentes.
* **Cantidad**: la cantidad de datos debe ser suficiente para entrenar un modelo efectivo.
* **Diversidad**: los datos deben ser lo suficientemente diversos para que el modelo generalice bien.
* **Etiquetado**: las etiquetas o respuestas asociadas deben ser precisas y consistentes.

**Figura 5**



*Nota*. Tipos de datasets [(ArcGIS, n.d.)](#rzxyimq79zgi)

### 

### Plataforma HuggingFace

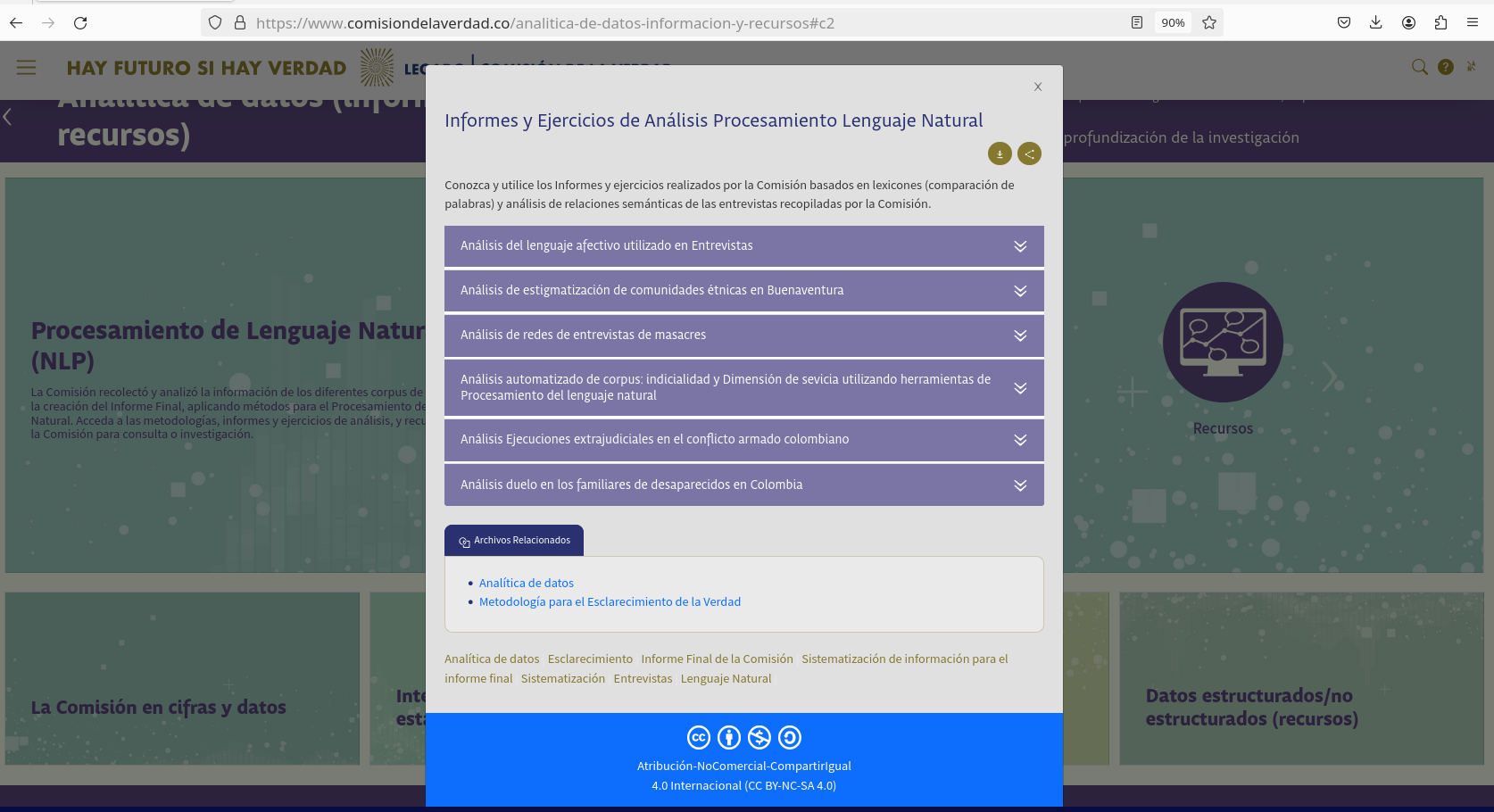
HuggingFace es una plataforma y comunidad de aprendizaje automático que proporciona herramientas y recursos para desarrollar, entrenar, compartir y desplegar modelos de inteligencia artificial. Es conocida por su biblioteca Transformers con soporte para múltiples lenguajes de programación, que ofrece una amplia gama de modelos pre-entrenados y herramientas para tareas de PLN como traducción, resumen, clasificación de texto, y más. HuggingFace proporciona la infraestructura necesaria para desplegar modelos de manera automática. Usando su API Inference, los modelos pueden ser desplegados en servidores gestionados por HuggingFace, permitiendo que se realicen inferencias en tiempo real. (HuggingFace, n.d.)

## Antecedentes

### Procesamiento del Lenguaje Natural Por Parte de La Comisión de La Verdad

La Comisión de la Verdad, establecida como parte del Acuerdo de Paz de 2016, tiene como finalidad esclarecer los hechos y promover una paz duradera en el país. Según el Decreto 588 de 2017, artículo 2º, uno de sus objetivos es aclarar lo sucedido (Legislación Colombiana, n.d.). Por ello, una de sus tareas más arduas ha sido la recopilación de datos, mayormente en lenguaje natural.

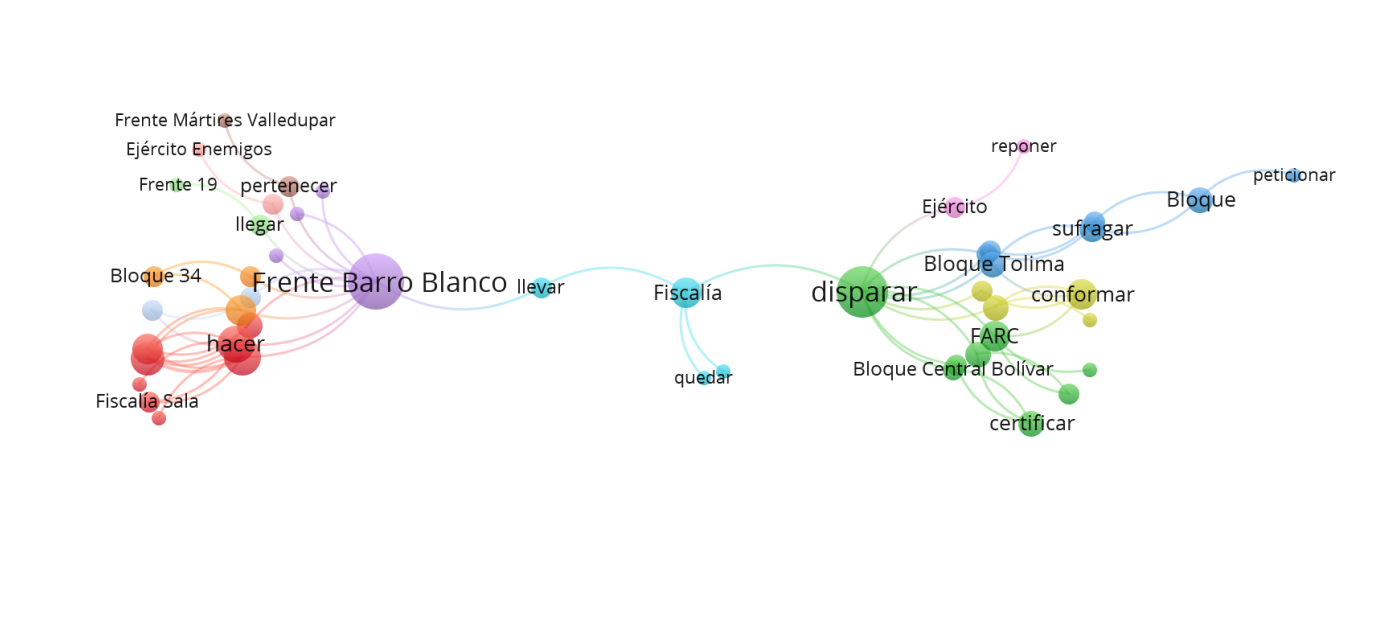
**Figura 6**



*Nota*. Proyectos con PLN de la Comisión de la verdad [(Comisión De La Verdad, 2022)](#jld8vix4fln).

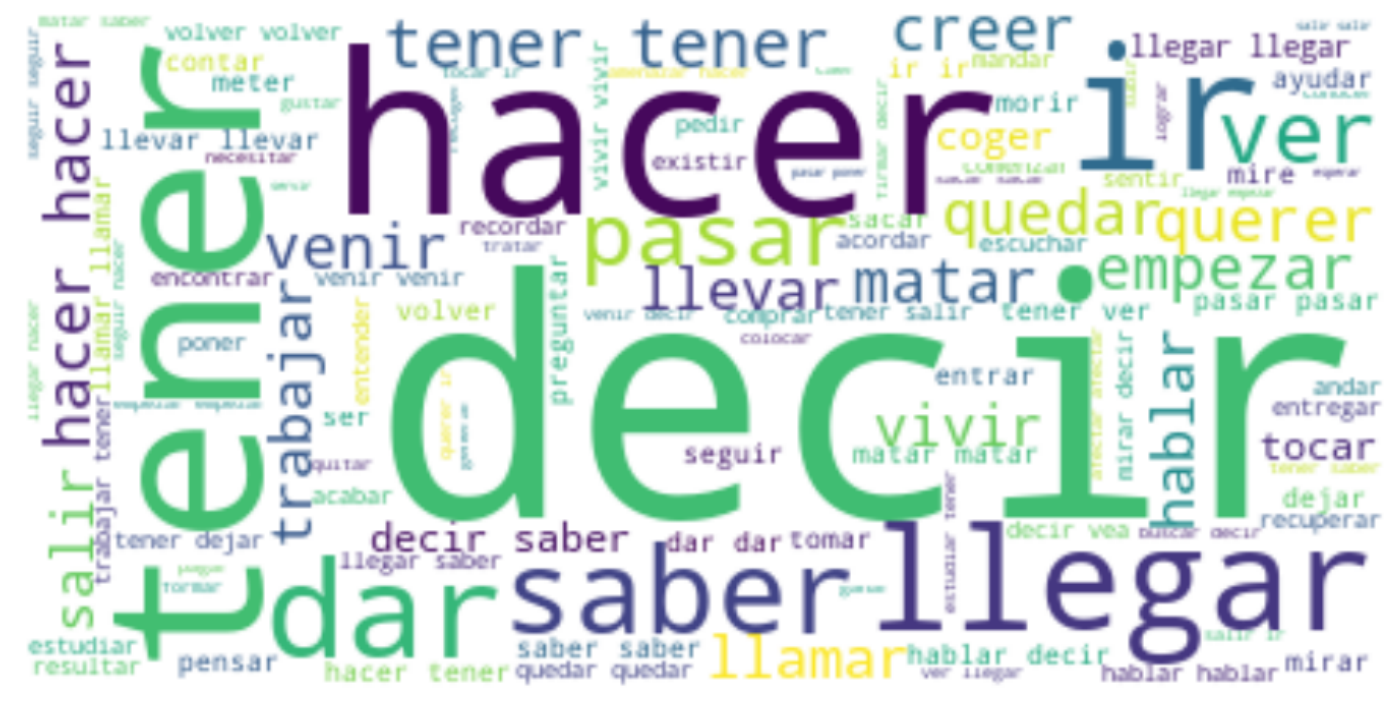
Entre los proyectos más destacados se encuentra Análisis automatizado de corpus: Se utilizaron herramientas de PLN para analizar grandes volúmenes de texto, como informes oficiales, testimonios y noticias, para identificar patrones e indicadores de violencia durante el conflicto. Esto permitió a la Comisión obtener una comprensión más profunda de la naturaleza del conflicto y sus efectos en la población.

**Figura 7**



*Nota*. Organizaciones-Verbos [(Comisión De La Verdad, 2022, p. 12)](#eza55fg8ydp).

**Figura 8**

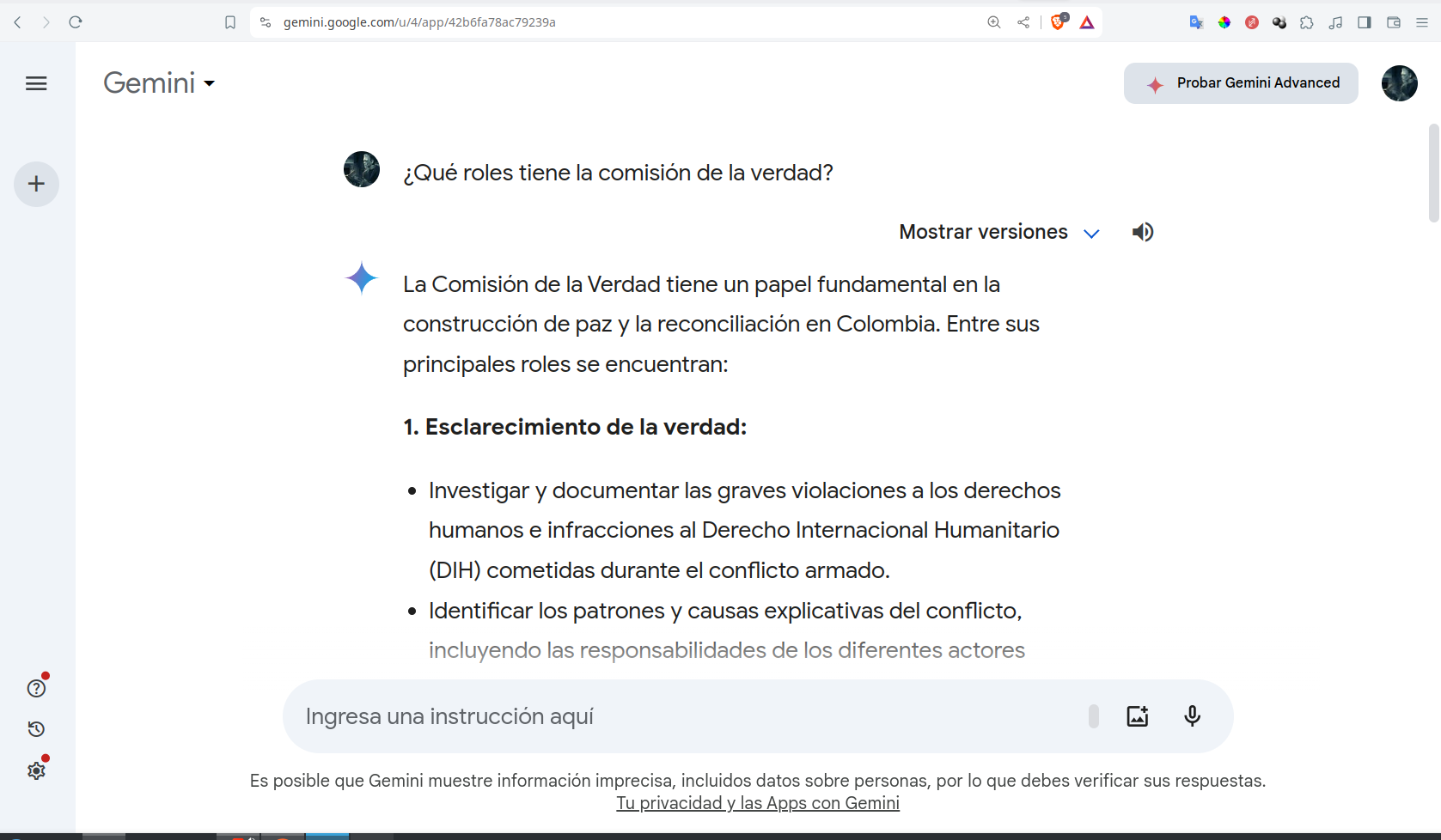


*Nota*. Nube de palabras de los 50 verbos más frecuentes. [(Comisión De La Verdad, 2022, p. 89)](#eza55fg8ydp).

### Google Gemini - Chatbot con Datos de la web

Gemini es un modelo de lenguaje grande (LLM) desarrollado por Google AI, entrenado en un conjunto de datos masivo de texto y código. Es capaz de realizar una amplia gama de tareas, incluyendo consultar información de cualquier tema. A Gemini se le pueden hacer preguntas generales sobre la comisión de la verdad. Aunque no tiene conocimiento específico de cada uno de los informes.

**Figura 9**

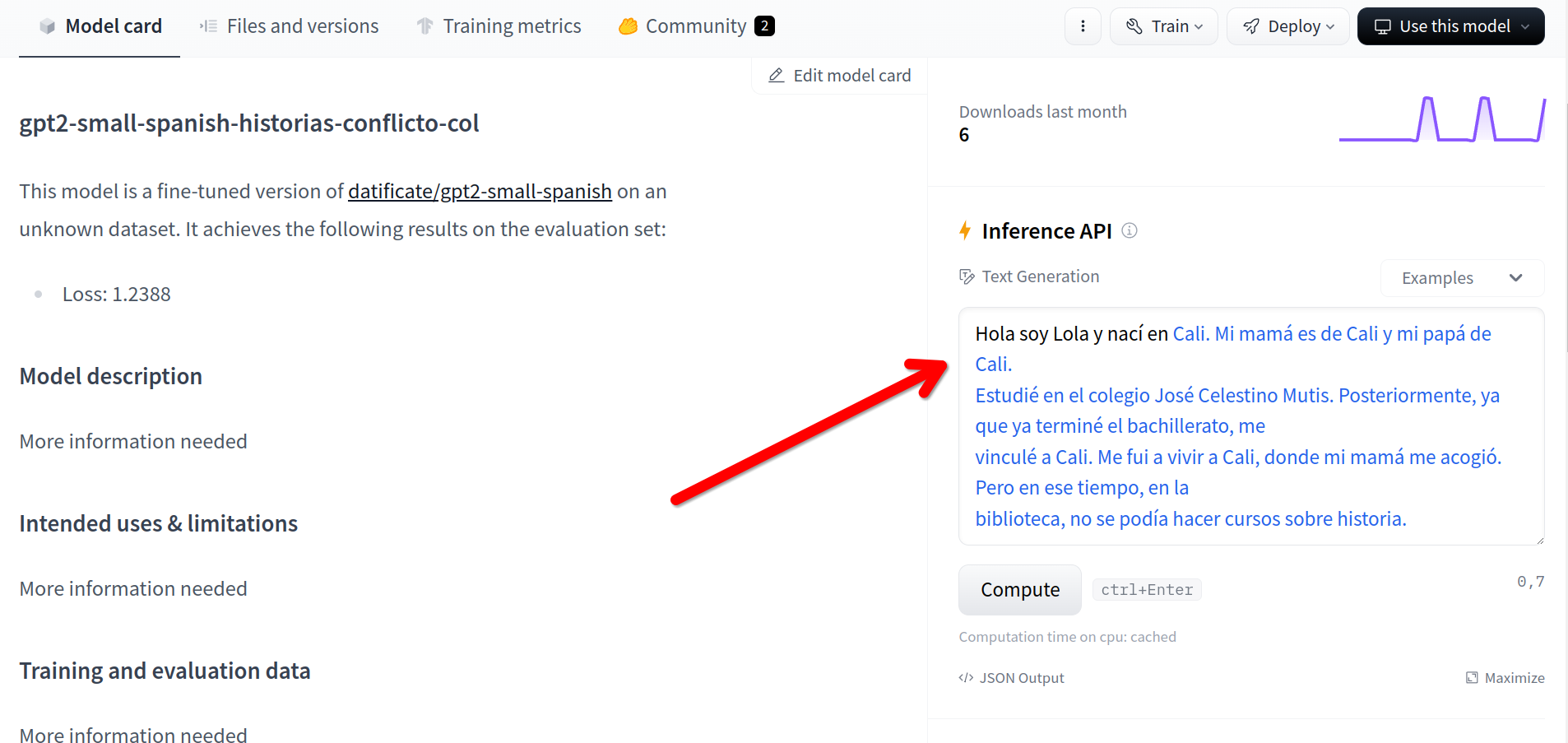


*Nota*. Gemini contestando preguntas sobre La Comisión de La Verdad [(Google AI, 2023)](#3kc50ogr6r20).

### Modelo Entrenado Con Historias Del Conflicto

Un modelo para generador de texto entrenado con testimonios del conflicto armado colombiano. Se partió del modelo [GPT-small-spanish](https://huggingface.co/datificate/gpt2-small-spanish) al cual se le hizo fine-tunning con un dataset etiquetado de 456 ejemplos de historias del conflicto armado en Colombia. El modelo genera la siguiente frase a partir de una frase inicial.

**Figura 10**

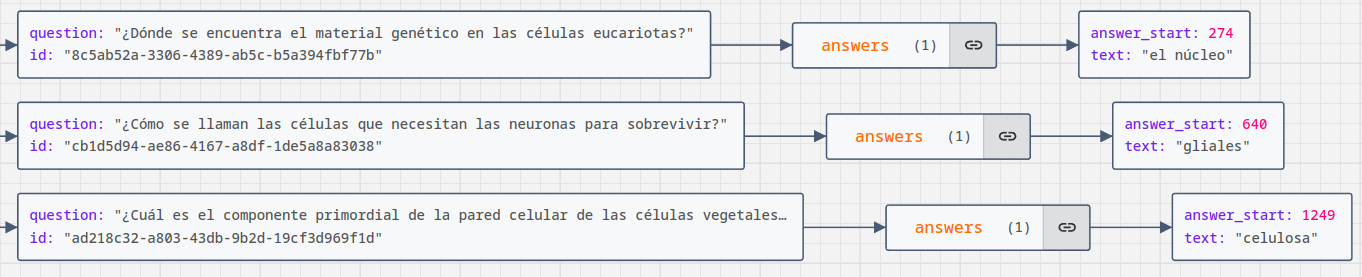


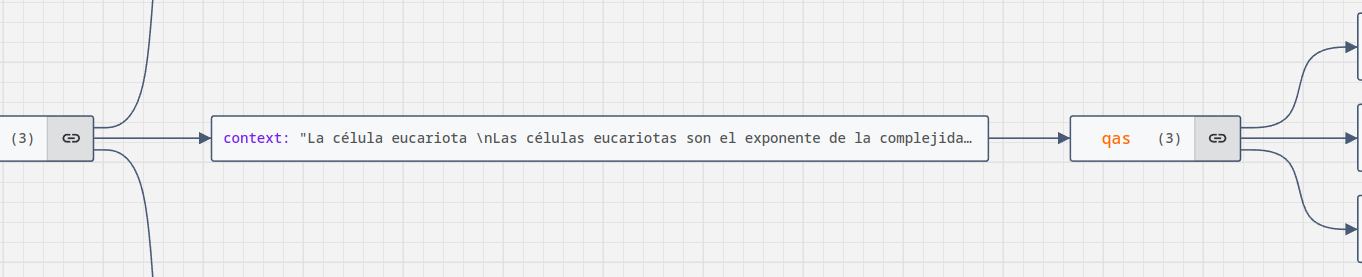
*Nota*. Modelo generativo gpt2-small-spanish-historias-conflicto-col [(Henao, 2022)](#9h2sigqqhvn4).

### Dataset de Pregunta-Respuesta en Español

El dataset SQAC de Plan de Tecnologías del Lenguaje - Gobierno de España (PlanTL-GOB) es un conjunto de datos en español diseñado para evaluar la comprensión de lectura y la capacidad de responder preguntas. Este dataset contiene 6.247 contextos y 18.817 preguntas con sus respectivas respuestas, variando de 1 a 5 preguntas por contexto. Los contextos provienen de diversas fuentes como artículos enciclopédicos de la Wikipedia en español y artículos de noticias de Wikinews [(PlanTL-GOB, 2023).](#i6cd8wc7xkxk)

**Figura 11**





Nota: Ejemplo del dataset SQAC. Elaboración propia.

### Conclusión Antecedentes

A partir de la revisión de antecedentes, se evidencia que si bien existen diversos proyectos que han aplicado técnicas de procesamiento del lenguaje natural (PLN) a los informes de la Comisión de la Verdad, este trabajo de grado se distingue por las siguientes características:

* Enfoque en la construcción de un sistema QA: A diferencia de los proyectos anteriores que se han centrado en el análisis de grandes volúmenes de texto, este trabajo de grado se propone desarrollar un dataset anotado y un sistema QA que permita a los usuarios acceder a información específica de manera rápida y eficiente.
* Utilización de un LLM de última generación: El sistema QA se basará en grandes modelos de lenguaje (LLM) como LLama, lo que permitirá ofrecer una mayor precisión y comprensión del lenguaje natural.
* Fine-tuning del modelo LLM: El LLM se afinará con un conjunto de datos de preguntas y respuestas sobre el conflicto armado colombiano, lo que permitirá mejorar su rendimiento en esta tarea específica.

En conjunto, este trabajo de grado tiene el potencial de hacer una contribución significativa al facilitar el acceso a la información contenida en los informes de la Comisión de la Verdad, lo que puede contribuir a la investigación, el análisis y la comprensión del conflicto armado colombiano y el proceso de paz.

# **Metodología**

## Actividades

Tabla 2

## *Actividad relacionada con cada objetivo.*

| **Objetivo Específico** | **Resultado Esperado por Actividad** | **Actividad** |
| --- | --- | --- |
| 1. Desarrollar un dataset anotado para QA | 1.1 Recopilación de informes de La Comisión de la Verdad. | 1. Búsqueda y descarga de informes en formato digital. |
|  | 1.2 Anotación de preguntas y respuestas. | 2. Definición de una guía de anotación. |
|  |  | 3. Realización del proceso de anotación. |
|  | 1.3 Documentación del dataset. | 4. Creación de un archivo de metadatos que describa el dataset. |
| 2. Implementar técnicas de fine-tuning con los informes de La Comisión De La Verdad y el dataset anotado usando uno o varios LLM | 2.1 Selección de uno o varios LLMs que sean adecuados. | 5. Investigar y revisar difere ntes LLM disponibles. |
|  | 2.2 Preparación del dataset para fine-tuning. | 6. División del dataset en conjuntos de entrenamiento, validación y prueba. |
|  | 2.3 Entrenamiento del modelo. | 7. Configuración de los parámetros de entrenamiento. |
|  |  | 8. Ejecución del entrenamiento del modelo. |
|  | 2.4 Evaluación del modelo. | 9. Evaluar respuestas sobre un conjunto de validación. |
|  | 2.5 Selección del modelo final. | 10. Comparación del rendimiento de diferentes modelos entrenados. |
| 3. Integrar el modelo afinado a la plataforma HuggingFace | 3.1 Un modelo alojado en HuggingFace | 11. Implementar las configuraciones necesarias para publicar el modelo. |
|  | 3.2 Un pipeline de HuggingFace utilizable desde la interfaz de la plataforma. | 12. Implementar las configuraciones necesarias para crear y publicar un pipeline. |
| 4. Testear el sistema QA | 4.1 Definición de casos de prueba. | 13. Identificación de diferentes tipos de preguntas y escenarios. |
|  | 4.2 Pruebas y validación de la pipeline. | 14. Probar con diversos ejemplos el pipeline de HuggingFace. |
|  | 4.3 Ejecución de pruebas. | 15. Utilización de herramientas de automatización de pruebas si es posible. |
|  | 4.4 Un informe con análisis de resultados y reconocimiento de errores. | 16. Identificación de errores y problemas en el sistema. |
|  |  | 17. Implementación de correcciones y mejoras en el sistema. |

## 

## Cronograma de Actividades

# 

**Tabla 3**

Cronograma de ocho meses.

# 

| ***Actividad*** | ***Tiempo en Meses*** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***1*** | ***2*** | ***3*** | ***4*** | ***5*** | ***6*** | ***7*** | ***8*** |
| 1 |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |  |  |
| 6 |  |  |  |  |  |  |  |  |
| 7 |  |  |  |  |  |  |  |  |
| 8 |  |  |  |  |  |  |  |  |
| 9 |  |  |  |  |  |  |  |  |
| 10 |  |  |  |  |  |  |  |  |
| 11 |  |  |  |  |  |  |  |  |
| 12 |  |  |  |  |  |  |  |  |
| 13 |  |  |  |  |  |  |  |  |
| 14 |  |  |  |  |  |  |  |  |
| 15 |  |  |  |  |  |  |  |  |
| 16 |  |  |  |  |  |  |  |  |
| 17 |  |  |  |  |  |  |  |  |

## 

## 

## 

## Presupuesto

**Tabla 4**

| Rubro | Estudiante | Universidad | Total |
| --- | --- | --- | --- |
| Recursos Humanos | $6,400,000 | $5,120,000 | $11,520,000 |
| Juan David Tovar | $6,400,000 | $0 |  |
| Raul E. Gutierrez de Piñerez Reyes |  | $5,120,000 |  |
| Hardware | $2,000,000 | $0 | $2,000,000 |
| Computador | $2,000,000 | $0 |  |
| Software y Otros servicios | $2,400,000 | $0 | $2,400,000 |
| Internet | $800,000 | $0 |  |
| Google Colab | $1,600,000 | $0 |  |
| Total | $10,800,000 |  | $15,920,000 |

# **Bibliografía**

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544-551.

ArcGIS. (n.d.). *Datasets de entidades en ArcGIS*. Datasets de entidades en ArcGIS Pro—ArcGIS Pro | Documentación. Retrieved May 27, 2024, from https://pro.arcgis.com/es/pro-app/latest/help/data/feature-datasets/feature-datasets-in-arcgis-pro.htm

Brazdil, P., van Rijn, J. N., Soares, C., & Vanschoren, J. (2022). *Metalearning: Applications to Automated Machine Learning and Data Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-030-67024-5\_4

Caballero, M. (2021). A Brief Survey of Question Answering Systems. *International Journal of Artificial Intelligence & Applications*, *Vol.12*(5). https://ssrn.com/abstract=3996229

Chase, H. (2023, 04). *"document question-answering" applications is a retrieval step,*. x.com. Retrieved May 27, 2024, from https://x.com/hwchase17/status/1649428297816711174/photo/1

Comisión De La Verdad. (2017, Abril 5). *Pagina Principal - Comisión De La Verdad*. Inicio | Informe Final Comisión de la Verdad. Retrieved April 19, 2024, from https://www.comisiondelaverdad.co/

Comisión De La Verdad. (2022). *Analítica de datos (información y recursos) | Informe Final Comisión de la Verdad*. Anexo. Retrieved May 16, 2024, from https://www.comisiondelaverdad.co/analitica-de-datos-informacion-y-recursos#c2

Comisión De La Verdad. (2022). *Metodolog´ıas de an´alisis de entrevistas utilizando Procesamiento de Lenguaje Natural (PLN)*. Comisión. Retrieved May 27, 2024, from <https://www.comisiondelaverdad.co/sites/default/files/descargables/2022-08/03_Anexo_NLP_07312022_4%2C9KB.pdf>

Comisión de la Verdad. (2022). Hay futuro si hay verdad | Informe Final Comisión de la Verdad. Comisión de la Verdad. Retrieved June 26, 2024, from https://www.comisiondelaverdad.co/hay-futuro-si-hay-verdad

DANE. (2017). *Encuesta Nacional de Lectura - ENLEC 2017*. DANE Información Para Todos. Retrieved 06 21, 24, from https://sitios.dane.gov.co/enlec\_dashboard/#!/

Deepset.ai. (2023, September 29). *Fine-tuning large language models*. Deepset. Retrieved May 27, 2024, from https://www.deepset.ai/blog/llm-finetuning

GIRALDO HOYOS, D. A. (2016, 7 15). *ASISTENTE VIRTUAL CHATBOT PARA LA WEB DE LA UNIVERSIDAD DEL VALLE*. https://acortar.link/I8kWsd

Google AI. (2023). *Google Gemini*. ‎Gemini - chat to supercharge your ideas. Retrieved May 16, 2024, from https://gemini.google.com/app

Hardesty, L. (2017, April 14). *Explained: Neural networks*. MIT News. Retrieved May 27, 2024, from https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

Henao, J. (2022). *gpt2-small-spanish-historias-conflicto-col*. huggingface.co. Retrieved May 27, 2024, from https://huggingface.co/jorge-henao/gpt2-small-spanish-historias-conflicto-col

H. Renear, A., Sacchi, S., & M. Wickett, K. (2011, 02). Definitions of Datasetin the Scientific and Technical Literature. <https://doi.org/10.1002/meet.14504701240>

HuggingFace. (n.d.). *Hugging Face – The AI community building the future.* The collaboration platform. Retrieved June 25, 2024, from https://huggingface.co/

Ji, Y., Wang, Q., & Li, X. (2019). A Survey on Tensor Techniques and Applications in Machine Learning. *IEEE Access*, *7*. 10.1109/ACCESS.2019.2949814

Kaur, P., Siddharth Kashyap, G., Kumar, A., Nafis, T., Kumar, S., & Shokeen, V. (2024). From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility. https://doi.org/10.48550/arXiv.2402.16142

La Web del Programador. (n.d.). *DICCIONARIO INFORMÁTICO*. Retrieved April 19, 2024, from https://www.lawebdelprogramador.com/

Legislación Colombiana. (n.d.). *DECRETO 588 DE 2017*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=80633

Lin, X., Wang, W., & Li, Y. (2024, Enero 30). Data-efficient Fine-tuning for LLM-based Recommendation. *arXiv - Cornell University*. https://arxiv.org/abs/2401.17197

Minaee, S., & Mikolov, T. (2024, Feb 20). Large Language Models: A Survey. *arXiv*.

PlanTL-GOB. (2023, October 12). *PlanTL-GOB-ES/roberta-large-bne-sqac · Hugging Face*. Hugging Face. Retrieved May 27, 2024, from https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne-sqac

Quinn, J., McEachen, J., Fullan, M., Gardner, M., & Drummy, M. (2019). *Dive Into Deep Learning: Tools for Engagement*. SAGE Publications.

Universidad De Los Andes. (2024). *Datasets de Administración*. Datasets de administración - Facultad de Administración. Retrieved May 27, 2024, from https://uniandes-co.libguides.com/administracion/datasets

Vaswani, A., Shazeer, N., & Parmar, N. (2017). Attention Is All You Need. *ArXiv*. ArXiv. 10.48550/arXiv.1706.03762 Focus to learn more.

Xu, R., Feng, Y., & Chen, H. (2023). ChatGPT vs. Google: A Comparative Study of Search Performance and User Experience. doi:10.48550/ARXIV.2307.01135