UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Modelos de Lenguaje Grande para el Análisis y Verificación de Hechos en Discursos Políticos

Proyecto de Titulación

Brayan Alexis Lechón Andrango

Alejandro Proaño Lozada, Ph.D.

Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito para la obtención del título de Magíster en Inteligencia Artificial

Quito, 18 de diciembre de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ
COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Modelos de Lenguaje Grande para el Análisis y Verificación de Hechos en Discursos Políticos

Brayan Alexis Lechón Andrango

Nombre del Director del Programa:	Felipe Grijalva
Título académico:	Ph.D. en Ingeniería Eléctrica
Director del programa de:	Inteligencia Artificial
Nombre del Decano del colegio Académico:	Eduardo Alba
Título académico:	Doctor en Ciencias Matemáticas
Decano del Colegio:	Ciencias e Ingenierías
Nombre del Decano del Colegio de Posgrados:	Dario Niebieskikwiat
Título académico:	Doctor en Física

Quito, 18 de diciembre 2024

© DERECHOS DE AUTOR

Nombre del estudiante:	Brayan Alexis Lechón Andrango
Código de estudiante:	00338860
C.I.:	1724676927
Lugar y fecha:	Quito, 18 de Diciembre de 2024

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following graduation project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

DEDICATORIA

A mi familia, quienes con su apoyo incondicional me inspiraron a seguir creciendo académicamente; siempre serán mi principal motivación, este logro no habría sido posible sin su eterna confianza en mí. A mis amigos, aquellos que estuvieron conmigo antes de iniciar este camino y también a quienes se unieron durante el proceso. Gracias por compartir las alegrías y los desafíos de este viaje. Finalmente, dedico este logro a mi abuela María Esther, quien lamentablemente partió de este mundo este año. Este trabajo es un testimonio del impacto que tuvo en mi vida, un legado que perdurará en el tiempo.

AGRADECIMIENTOS

En primer lugar, quiero expresar mi más sincero agradecimiento a Alejandro Proaño por su guía y compromiso durante el desarrollo de este trabajo. Su conocimiento fue fundamental para la culminación de este proyecto. Agradezco también a todos los docentes de la USFQ, quienes, con su dedicación y compromiso con la educación, contribuyeron de manera significativa a mi crecimiento académico y profesional. Por último, extiendo mi gratitud a todas las personas que aportaron, de manera directa o indirecta, al desarrollo de este trabajo. Su apoyo fue esencial para alcanzar esta meta.

RESUMEN

Este estudio explora la aplicación de Modelos de Lenguaje Grande (LLMs) para el análisis de discursos políticos y la verificación de hechos en Ecuador. La investigacion cubre tres tareas principales: la generación de resúmenes concisos, el análisis de sentimiento y la verificación de hechos utilizando un enfoque híbrido que combina fine-tuning y Retrieval-Augmented Generation (RAG). Los resultados experimentales demuestran la efectividad de los LLMs en la gestión de validación de datos complejos, contribuyendo a una mayor transparencia en el discurso político ecuatoriano.

Palabras clave: LLMs, Retrieval-Augmented Generation, Fine-Tuning, Verificación de Hechos, Discursos Políticos

ABSTRACT

This study explores the application of Large Language Models (LLMs) for political speech analysis and fact-checking in Ecuador. We investigate three main tasks: generating concise summaries, conducting sentiment analysis, and fact-checking using a hybrid approach that combines fine-tuning and Retrieval-Augmented Generation (RAG).Experimental results demonstrate the LLM's effectiveness in managing complex data validation, contributing to enhanced transparency in Ecuadorian political discourse.

Key words: LLMs, Retrieval-Augmented Generation, Fine-Tuning, fact-Check, Political Speeches

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Modelos de Lenguaje Grande para el Análisis y Verificación de Hechos en Discursos Políticos

Brayan Alexis Lechón, *Member, IEEE,*

***Resumen*—Este estudio explora la aplicación de Modelos de Lenguaje Grande (LLMs) para el análisis de discursos políticos y la verificación de hechos en Ecuador. La investigacion cubre tres tareas principales: la generación de resúmenes concisos, el análisis de sentimiento y la verificación de hechos utilizando un enfoque híbrido que combina fine-tuning y Retrieval-Augmented Generation (RAG). Los resultados experimentales demuestran la efectividad de los LLMs en la gestión de validación de datos complejos, contribuyendo a una mayor transparencia en el discurso político ecuatoriano.**

***Index Terms*—LLMs, Retrieval-Augmented Generation, Fine-Tuning, Verificación de Hechos, Discursos Políticos.**

I. Introducción

DESDE los inicios del estudio de la inteligencia artificial, el procesamiento del lenguaje natural (NLP, por sus siglas en inglés) ha sido un área de gran interés, evolucionando de sistemas basados en reglas a enfoques estadísticos y modelos de aprendizaje profundo. Con el reciente desarrollo de los modelos de atención, este campo ha experimentado un avance notable, permitiendo la creación de Modelos de Lenguaje Grande (LLMs, por sus siglas en inglés) que han marcado una nueva era en la interacción entre humanos y máquinas. El crecimiento constante de datos en formato textual, combinado con los avances en capacidad computacional, ha acelerado la innovación en diversas aplicaciones, tales como sistemas de recomendación, motores de búsqueda, chatbots y asistentes virtuales. Además, la democratización de estos modelos por parte de empresas como OpenAI, Meta y Anthropic ha ampliado las posibilidades de crear herramientas avanzadas basadas en el procesamiento del lenguaje natural.

A pesar de los avances significativos en los LLMs, estos presentan limitaciones. Los LLMs pueden generar contenido falso, fabricado o inconsistente con sus datos de entrenamiento, un fenómeno conocido como *alucinaciones*. Este comportamiento puede minimizarse en cierta medida, pero estudios muestran que las alucinaciones en modelos de lenguaje son inevitables debido a sus limitaciones estructurales inherentes [1]. Además, otro desafío importante es su incapacidad para integrar nueva información, para ello se debe realizar un reentrenamiento que suele ser costoso. Estas limitaciones han impulsado el desarrollo de técnicas que buscan personalizar y adaptar los modelos para mejorar su precisión en entornos donde se requiere información actualizada. Un enfoque muy popular es *ReAct* [2], el cual permite a los modelos de lenguaje generar acciones específicas de manera intercalada, facilitando así el manejo de información dinámica.

Finalmente, el aspecto ético también es crucial. Se ha demostrado que los LLMs pueden tener sesgos derivados de la falta de datos en determinadas áreas, lo que afecta la validez de las respuestas generadas. Estos sesgos, junto con la alta confianza de los usuarios en las respuestas de los modelos (especialmente en áreas fuera de su propio campo de expertise), subrayan la necesidad de seguir trabajando en la mejora y validación de los LLMs antes de su implementación en aplicaciones.

A. Motivación

El crecimiento del contenido en formato de texto, impulsado por redes sociales y otras plataformas digitales, ha dificultado cada vez más la distinción entre noticias falsas y hechos verificables, especialmente en el ámbito político. En Ecuador, un estudio evidenció que en las elecciones de 2021 el 50 % de los votantes emitieron su voto influenciados por información incorrecta divulgada en redes sociales como Facebook, afectando su capacidad de discernir entre noticias falsas y reales [3].

Además, el creciente uso de LLMs en la generación automatizada de contenido político, como campañas publicitarias y discursos, agrava este problema al facilitar la propagación de información manipulada [4]. Esto afecta tanto la percepción pública como la integridad de los procesos electorales, representando un desafío considerable para la política moderna. Verificar los hechos en discursos políticos se ha vuelto una tarea compleja y crítica, ya que la información inexacta puede ser fácilmente manipulada para influir en la opinión ciudadana sobre los candidatos en distintos niveles políticos.

Este artículo presenta un caso de estudio en Ecuador que utiliza LLMs para analizar y verificar información en discursos políticos, con el objetivo de establecer un método confiable para distinguir hechos verificables de posibles manipulaciones. A través de esta investigación, se explora el potencial de los LLMs como herramientas para mejorar la transparencia y la credibilidad en la comunicación política, sentando las bases para futuras aplicaciones en contextos políticos complejos.

B. Lechón esta con la Universidad San Francisco de Quito, Quito, Ecuador, e-mail: blechon@estud.usfq.edu.ec

II. Estado del Arte

En este trabajo consideramos dos areas de investigación. En primer lugar, se examinan los métodos actuales empleados para el ajuste de los Modelos de Lenguaje de Grande (LLMs, por sus siglas en inglés) a temáticas o tópicos específicos. Segundo abordaremos las aplicaciónes de estos modelos en el análisis de discursos políticos y en la validación de hechos.

A. Métodos para Adaptar un LLM

La personalización de los LLMs es fundamental para optimizar su rendimiento en tareas específicas sin requerir el desarrollo desde cero, que suele demandar enormes cantidades de datos y recursos computacionales [5]. Diversas técnicas han surgido para facilitar esta adaptación en función del poder computacional disponible y los objetivos de la aplicación a desarrollar.

La Figura 1 presenta un resumen de los principales métodos de ajuste y entrenamiento empleados actualmente en LLMs, ilustrando la relación entre el rendimiento del modelo y la complejidad o esfuerzo computacional requerido para cada técnica.

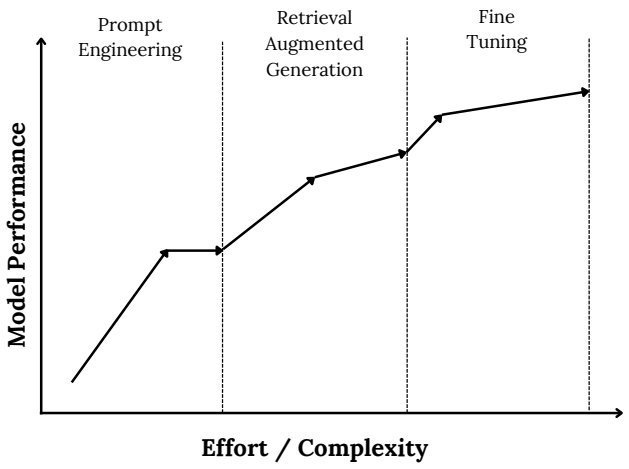

Figura 1. Técnicas para entrenar un LLM

1). Prompt Engineering: La ingeniería de prompts (*prompt engineering*) se centra en diseñar instrucciones específicas para optimizar el rendimiento sin modificar el modelo. El diseño debe ser meticuloso; como muestran Liu et al. [6], pequeños cambios en el prompt pueden llevar a variaciones significativas en los resultados.

Diversos enfoques de prompt engineering, como *Role-Prompting*, *Chain-of-Thought*, *ReAct*, entre otros, han demostrado mejorar el rendimiento del modelo en tareas variadas [7]. La elección del método adecuado es crucial para alinearse con las demandas específicas de cada aplicación.
2). Retrieval-Augmented Generation (RAG): La Generación Aumentada por Recuperación (RAG) extiende las capacidades de los LLMs mediante la recuperación de información relevante de fuentes externas [8]. RAG extrae información relevante basada en una relación semántica con el prompt proporcionado por el usuario. Como describen Gao et al. [9], este método ha demostrado ser particularmente útil en aplicaciones de preguntas y respuestas, proporcionando respuestas verificables y específicas, incluso en dominios abiertos.
3). Fine-Tuning: El *fine-tuning* consiste en adaptar un LLM preentrenado mediante datos específicos, mejorando su desempeño en tareas especializadas. Aunque produce buenos resultados, requiere grandes volúmenes de datos y puede llevar a problemas como sobreajuste o pérdida de conocimientos previos. Para abordar estas limitaciones, técnicas como LoRA y QLoRA [10], [11] permiten ajustar solo módulos específicos del modelo, congelando otros para reducir el consumo de memoria y mantener el rendimiento.

B. LLMs en aplicaciones Políticas

El uso de LLMs en el contexto político es un área de gran interés. En su estudio, Rettenberger et al. [12] analizaron los sesgos presentes en modelos de lenguaje como LLaMA-70B, encontrando que este modelo muestra un sesgo hacia posiciones liberales, especialmente cuando se utiliza en inglés. Sin embargo, estas inclinaciones pueden variar según el idioma; en otros idiomas, como el alemán, los modelos revelan una polarización más marcada. Esta variabilidad podría atribuirse a la falta de datos equilibrados en diferentes idiomas y contextos políticos, lo que limita la capacidad del modelo para representar de manera adecuada la diversidad de opiniones.

La opinión pública no siempre se encuentra alineada con el contenido generado por los LLMs, especialmente en temas políticos. Santurkar et al. [13] observaron que los LLMs reflejan sesgos que reproducen las divisiones políticas actuales, tendiendo hacia una inclinación liberal y una representación limitada de la diversidad de opiniones, especialmente de grupos minoritarios o subrepresentados. El uso de LLMs en este ámbito requiere un desarrollo ético; es fundamental combinar múltiples enfoques para mejorar la representatividad y reducir los sesgos.

C. Validación de Hechos con LLMs

El uso de modelos de lenguaje de grande para la validación de hechos presenta una gran oportunidad si se gestiona adecuadamente. Modelos como GPT-3 y LLaMA han demostrado ser poderosos en la identificación de desinformación; sin embargo, también pueden ser un arma de doble filo. Chen et al. [14] destacan que, debido a la falta de acceso a información actualizada y a sesgos en los datos de entrenamiento, estos modelos son propensos a generar alucinaciones o errores factuales, complicando la detección precisa de desinformación en el contenido generado.

La validación de hechos con LLMs ha sido abordada por diversos estudios utilizando enfoques que, en su mayoría, no requieren ajuste fino. Por ejemplo, Li et al. [15] desarrollaron Self-Checker, un sistema de módulos basados en *prompting* que mejora la precisión de los modelos sin

necesidad de modificar su estructura original. Zhang et al. [16] proponen un método jerárquico de *prompting* paso a paso para la verificación de hechos en noticias. Galitsky et al. [17] introducen una herramienta que utiliza algoritmos y procedimientos de verificación, trabajando en conjunto con LLMs para reducir sus alucinaciones.

El principal desafío consiste en evaluar con precisión la validez del conocimiento del modelo antes de recurrir a fuentes externas. La precisión de los modelos puede variar significativamente en función de la fuente de la información y su actualidad, lo que subraya la necesidad de combinar métodos complementarios para optimizar la exactitud en la verificación de hechos [18].

D. Enfoque Híbrido: Combinación de Retrieval y Fine-Tuning para la validación de Hechos

Tanto *RAG* como *fine-tuning* han demostrado ser técnicas altamente efectivas para optimizar el desempeño de los LLMs [19]. Ante los desafíos presentes en el análisis de discursos políticos y la validación de hechos, en este trabajo se propone un enfoque híbrido que integra las fortalezas del *fine-tuning*, que adapta el modelo a los matices y retóricas específicas del lenguaje político, con las ventajas de RAG, que proporciona acceso dinámico a fuentes de información actualizadas, esenciales para la verificación de hechos en tiempo real.

Este enfoque híbrido está emergiendo como una alternativa prometedora para la personalización de los LLMs, como se demuestra en estudios como *Retrieval-Augmented Dual Instruction Tuning* (RA-DIT) [20] y *RAFT* [21]. Estas investigaciones demuestran cómo el enfoque híbrido mejora la capacidad del modelo para identificar información relevante, reducir sesgos y alucinaciones, y ofrecer respuestas más equilibradas y contextuales.

III. Metodología

El análisis de discursos políticos se centra en tres módulos: resumen del texto, análisis de sentimiento y validación de hechos. Para cada tarea, se han desarrollado técnicas específicas, optimizadas para alcanzar resultados precisos y fiables. La Figura 2 ilustra el flujo de trabajo general implementado, estructurado en módulos independientes que garantizan un procesamiento eficiente.

A. Sumarización

El objetivo de este módulo es generar un resumen que sea conciso y coherente, capturando los puntos clave del discurso político mientras se preserva la intención y el tono del orador. Esto es especialmente crítico en discursos donde los matices del lenguaje juegan un papel importante.

1). Métodos Utilizados: Para esta tarea, se empleó el modelo GPT-4o de OpenAI mediante su API, seleccionado por su capacidad para procesar largas cantidades de información contextual [22]. El modelo procesa el discurso original mediante un *prompt* diseñado en base a *Role-Prompting* [23] en combinación con instrucciones secuenciales. Este enfoque se considera apropiado para orientar al modelo hacia la identificación de los aspectos más relevantes del discurso.

B. Análisis de Sentimientos

Este módulo busca identificar el tono emocional subyacente en el discurso político, clasificándolo como positivo, negativo o neutral. Dado que los discursos políticos suelen abarcar múltiples temas con diversos tonos emocionales, se procede a segmentar el texto en fragmentos que se denominan *ideas*, facilitando un análisis detallado y capturando así la evolución emocional del discurso.

1). Métodos Utilizados: Para la segmentación del discurso, se empleó GPT-4o, con un *prompt* estructurado de manera similar al del módulo de resumen. Este enfoque facilita la división del discurso en ideas clave, lo cual permite un análisis posterior más enfocado en las variaciones emocionales a lo largo del texto.

El análisis de sentimientos se llevó a cabo utilizando el modelo *distilbert-base-multilingual-cased-sentiments-student* [24], un modelo de BERT multilingüe optimizado para operar en modo *zero-shot*, lo cual resulta útil en escenarios donde no se dispone de datos etiquetados. Este enfoque asegura que cada fragmento se clasifique correctamente sin necesidad de un ajuste fino adicional, permitiendo obtener una evaluación detallada y generalizable del tono emocional en cada parte del discurso.

C. Validación

Para este módulo se busca implementar un sistema avanzado que pueda validar hechos presentes en los discursos políticos al estilo del fact-checking. Este módulo combina *fine-tuning* del modelo con RAG para ofrecer respuestas basadas en información reciente y verificable.

1). Conjunto de Datos: El conjunto de datos empleado para el *fine-tuning* se recopiló de plataformas especializadas en verificación de información como *Ecuador Verifica* [25] y el *Observatorio Legislativo* de la iniciativa Fundación Ciudadanía y Desarrollo [26]. Aunque estas plataformas se centran principalmente en la verificación de noticias falsas, se logró compilar 114 muestras relevantes sobre validación del discurso público que abordan temas como debates legislativos, discursos de campaña y eventos políticos clave como juicios políticos.

Las etiquetas se simplificaron de la siguiente forma para facilitar el entrenamiento:

- **Verdadero**: Información que ha sido comprobada y es completamente precisa.
- **Falso**: Información incorrecta o que no puede ser respaldada por fuentes oficiales.
- **Engañoso**: Afirmaciones que contienen elementos verdaderos, pero presentados de manera que distorsionan la realidad.

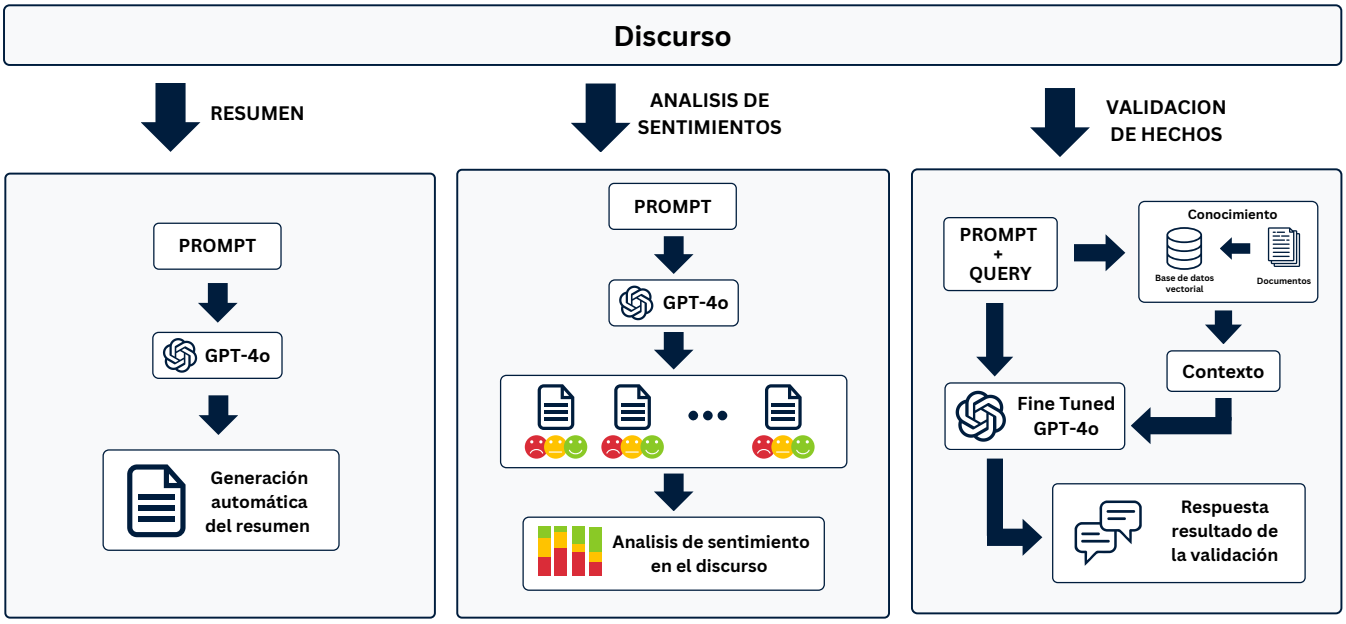

Figura 2. Proceso de análisis de discursos políticos con LLMs. El diagrama presenta una vista general del flujo de trabajo implementado, organizado en tres fases principales: resumen, análisis de sentimientos y validación de hechos.

2). Ajuste Fino (Fine-Tuning): El *fine-tuning* del modelo GPT-4o se llevó a cabo utilizando la API de OpenAI. Se estructuró el conjunto de datos en formato JSONL, con ejemplos detallados de la interacción esperada entre la pregunta de validación y sus respuestas correspondientes, conforme a las directrices que OpenAI especifica en su documentación [27]. Se utilizaron los siguientes hiperparámetros para optimizar el modelo:

- **Epochs**: 3
- **Batch size**: 1
- **Learning Rate Multiplier**: 1.8
- **Seed**: 13

Los resultados del entrenamiento se muestran en la Figura 3 donde se puede apreciar que la pérdida del entrenamiento inicialmente comienza en niveles altos (aproximadamente 2.5) y va disminuyendo hasta estabilizarse en valores cercanos a 0.5. Además, una vez terminado el entrenamiento, se evaluó el desempeño del modelo en términos de capacidad de respuesta en el Playground que ofrece OpenAI, donde se puede comparar dos modelos con los mismos inputs. Esto garantizó que el modelo afinado cumpliera con el comportamiento esperado. Gracias a este ajuste, el LLM se centra en validar la información como un fact-checker, generando un veredicto en su respuesta en base a las etiquetas que se establecieron previamente.

3). Generación Aumentada por Recuperación (RAG): Para mejorar aún más la precisión y relevancia de las respuestas, se implementó un sistema RAG que permite al modelo consultar información actualizada de manera dinámica. El proceso se desarrolló en tres etapas:

1. **Indexación de Documentos**: Se creó un índice vectorial basado en *embeddings*, utilizando *Pinecone* para un almacenamiento eficiente. La base vectorial se construyó mediante código específico, optimizando la rapidez en la búsqueda y recuperación.
2. **Recuperación de Información**: Frente a una afirmación a validar, el sistema emplea los *embeddings* para localizar y recuperar documentos relevantes. Esto incluye información crítica que puede respaldar o refutar la afirmación con base en datos verificados.
3. **Generación de Respuesta**: El modelo combina la información recuperada con su conocimiento interno, generando una respuesta detallada y bien fundamentada. Este enfoque híbrido garantiza una verificación

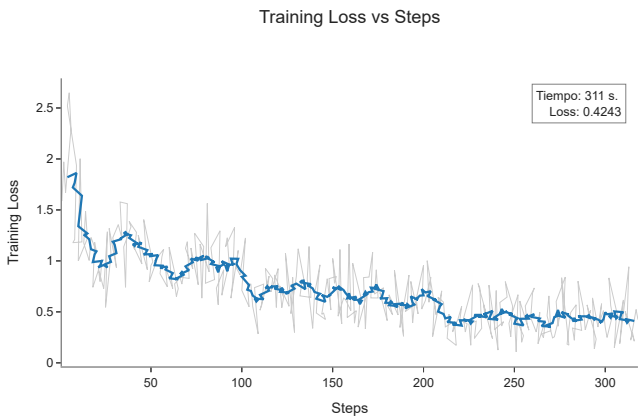

Figura 3. Curva de pérdida (loss) en función de los pasos de entrenamiento (steps) del modelo GPT-4o de OpenAI. La curva azul representa la tendencia de la pérdida promedio, mientras que las líneas grises muestran la variabilidad de la pérdida en cada paso.

precisa y contextualizada.

D. Base de Conocimiento Externa

La base de conocimiento externa, fundamental para el proceso RAG, se compone de documentos oficiales del Ecuador, como la Constitución, el Código Civil, el COIP y la Ley de Seguridad Pública y del Estado. Adicionalmente, se incorporaron boletines y estadísticas del INEC, informes macroeconómicos del Banco Central y reportes del Ministerio del Interior sobre temas de seguridad ciudadana, incluyendo homicidios, secuestros y uso de armas ilícitas.

Esta estructura permite que el sistema acceda a información rigurosamente verificada y actualizada hasta 2024, robusteciendo las capacidades del modelo para proporcionar respuestas precisas y relevantes en el contexto de la validación de hechos.

E. Interfaz de Usuario

Para facilitar la interacción, se desarrolló una interfaz intuitiva que permite al usuario ingresar el discurso junto con datos clave tales como el autor, el cargo (por ejemplo, asambleísta, ministro, presidente) y la fecha. Una vez ingresada esta información, el sistema analiza el discurso en las fases de resumen y análisis de sentimientos, generando un informe detallado. En la fase de validación de hechos, el usuario puede interactuar con el modelo en forma de chatbot, promoviendo una interacción dinámica y efectiva.

F. Experimentos

Los experimentos fueron diseñados para evaluar los módulos desarrollados en este trabajo: resumen, análisis de sentimiento y validación de hechos. Cada módulo fue probado utilizando un subconjunto específico del dataset descrito en la seccion de resultados.

1). Evaluación del Módulo de Resumen: La evaluación se centró en analizar la concisión y la retención de información clave, empleando las siguientes métricas:

- **Similitud Semántica**: Para medir qué tan bien el resumen captura la esencia del discurso original.
- **Reducción de Longitud**: Evaluamos la concisión del resumen generado mediante el cálculo del porcentaje de reducción de longitud respecto al discurso original.

2). Evaluación del Módulo de Análisis de Sentimiento: La evaluación de este módulo se centró en analizar la variabilidad emocional (positiva, negativa y neutra) en los discursos, con base en las ideas generadas por el modelo, para así identificar patrones y tendencias.

3). Evaluación de la Validación de Hechos: La evaluación de este módulo empleó un enfoque multidimensional, tal como explora Celikyilmaz et al. [28], se ha evidenciado que realizar evaluaciones en tres niveles —métricas centradas en el humano, métricas automáticas y métricas basadas en aprendizaje automático— ayuda a asegurar que el modelo no solo sea preciso en la clasificación, sino también relevante y coherente en la generación de sus respuestas. Las métricas empleadas en este estudio son las siguientes:

- **Matriz de Confusión**: Para medir la precisión en la clasificación de respuestas correctas e incorrectas.
- **Similitud Coseno**: Para evaluar la correspondencia entre las respuestas generadas y las validaciones oficiales.
- **Métricas de Calidad**: Para analizar la claridad, relevancia y fiabilidad de las respuestas generadas.

Para las métricas de calidad se diseñó una rúbrica cualitativa con cuatro criterios a analizar (Tabla I), de esta manera se busca evaluar la capacidad del sistema para manejar afirmaciones complejas y ofrecer un análisis detallado de su rendimiento más allá de la coincidencia de la etiqueta positiva.

Para que esta evaluación tenga un criterio mucho más matizado, el trabajo se apoyó en la metodología LLM as a Judge, que permite realizar un análisis en profundidad de las respuestas generadas por los modelos. Esta metodología, evaluada por Zheng et al. [29], se basa en emplear LLMs como evaluadores imparciales. De este modo, es posible valorar aspectos como la relevancia, precisión o profundidad, lo que enriquece significativamente el proceso de valoración y permite una comparación más rigurosa en las validaciones obtenidas. Para ello, se empleó el modelo ClaudeSonet3.5, que ha demostrado en algunos benchmarks estar a la altura, o incluso superar en ciertas tareas, al modelo GPT-4o [30].

TABLA I
MÉTRICAS DE EVALUACIÓN DE RESPUESTAS DE LA VALIDACIÓN DE HECHOS

Métrica	**Descripción**
Precisión del Veredicto	Evalúa si el veredicto proporcionado (Verdadero, Falso, Engañoso) es correcto y basado en evidencia verificable.
Claridad y Comprensibilidad	Mide si la explicación es clara, estructurada y fácil de seguir, facilitando la comprensión.
Relevancia y Solidez de la Evidencia	Examina la calidad y relevancia de la evidencia, verificando la fiabilidad de las fuentes y la justificación del veredicto.
Imparcialidad y Objetividad	Determina si la respuesta es imparcial, sin sesgos, y si presenta los hechos de manera equilibrada.

G. Dataset para los Experimentos

El conjunto de datos empleado en los experimentos se compone de transcripciones oficiales de diversas temáticas del ámbito político ecuatoriano. Estos datos se dividieron en subconjuntos específicos según los requerimientos de cada módulo experimental.

1). Discursos para Análisis de Resumen y Sentimiento: Para los módulos de resumen y análisis de sentimiento, se utilizó un subconjunto de 21 discursos realizados en debates de la asamblea nacional:

- Debate Ley de Turismo (Mayo 2024)

- Debate Ley de Seguridad Digital (Junio 2024)
- Debate Tratado de Libre Comercio Ecuador-China (Marzo 2024)
- Debate Presupuesto General del Estado (Marzo 2024)

Cada discurso fue procesado previamente para extraer el texto principal, eliminando redundancias y errores comunes en las transcripciones.

2). Validaciones para el Módulo de Verificación de Hechos: Para el módulo de validación de hechos, se empleó un subconjunto de 22 verificaciones que no formaron parte del conjunto de entrenamiento del modelo. Estas verificaciones se generaron en los siguientes escenarios:

- Informe de Rendición de Cuentas del Presidente Daniel Noboa (Mayo 2024)
- Juicio Político a la Ministra del Interior Mónica Palencia (Octubre 2024)
- Debates legislativos realizados en el pleno de la Asamblea Nacional (2024)

Las verificaciones fueron seleccionadas en base a la diversidad de temas abordados, prevaleciendo que las muestras contengan afirmaciones simples y complejas.

H. Repositorio del Proyecto

Todo el código, análisis y resultados encontrados en esta investigación se encuentran disponibles en el Repositorio de GitHub *Analizador de Discursos* [31]. El conjunto de datos no se encuentra en el repositorio, pero puede recrearse siguiendo los pasos especificados en el archivo README; además, está disponible bajo solicitud.

IV. Resultados y Discusión

En esta sección se presentan y analizan los resultados obtenidos, organizados en dos aspectos: el análisis del discurso y la validación de hechos. Asimismo, se discuten estos hallazgos en el contexto del análisis de discursos políticos, así como sus limitaciones y oportunidades para futuros trabajos.

A. Resumen del Discurso

La Tabla II resume los resultados de la reducción de longitud y la similitud semántica, evaluada mediante la distancia coseno. La longitud del texto, tanto para el discurso original como para el resumen, se calculó en función de la cantidad de tokens presentes, de acuerdo con el modelo GPT-4o.

$$\text{Similitud} = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

$$\text{Reducción (\%)} = \left(1 - \frac{\text{Tokens}_{\text{resumen}}}{\text{Tokens}_{\text{original}}}\right) \times 100$$

En cuanto a la similitud, se observa una mediana superior a 0.8, lo cual indica que el resumen generado captura en gran medida los aspectos más importantes del texto original. Con una desviación estándar de 0.04 respecto a su media de 0.81, se puede inferir que el modelo mantiene una representación consistente de la información clave del discurso, sin variaciones significativas en la similitud entre diferentes discursos.

TABLA II
Estadísticas de Longitud y Similitud del Resumen

Est.	Original (Tokens)	Resumen (Tokens)	Reducción (%)	Similitud Coseno
mín	541	228	39 %	0.69
máx	1677	449	84 %	0.86
$\bar{x}$	1062	315	67 %	0.81
σ	366	66	13 %	0.04
P_{50}	1071	315	71 %	0.82

Por otro lado, en la reducción de longitud, la mediana se sitúa alrededor de 0.71, lo que indica que, en promedio, el resumen logra una compresión significativa respecto al texto completo. Además, la desviación estándar de 13 % respecto a su media de 67 % sugiere que el modelo es capaz de condensar el contenido sin comprometer excesivamente la similitud con el texto original.

Si bien los resultados generales sugieren que el modelo es efectivo en esta tarea de resumen, también se pueden observar ciertas anomalías en la relación entre similitud y reducción, presentadas en la Figura 4. Se identifican algunos textos generados con alta reducción y baja similitud (cercanos a 0.7), lo cual podría reflejar casos en los que la compresión excesiva conduce a una pérdida de detalles importantes. Esta variabilidad sugiere que, aunque el modelo generalmente logra un buen equilibrio entre concisión y preservación de la información, sería útil ajustar el diseño del prompt para priorizar la retención de información clave en discursos largos, optimizando así la calidad del resumen en estos casos más complejos.

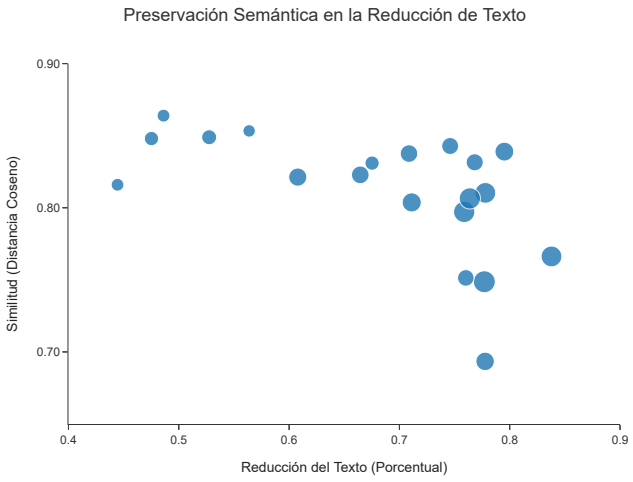

Figura 4. Relación entre Similitud y Reducción del Texto, considerando el tamaño del discurso. El tamaño de los puntos representa el número de tokens en el discurso original.

B. *Análisis de Sentimientos*

La Tabla III resume los resultados del análisis de sentimientos realizados en la muestra de discursos, mostrando las proporciones de sentimientos positivo, neutral y negativo, junto con el número de ideas identificadas en cada discurso.

TABLA III
ESTADÍSTICAS ANÁLISIS DE SENTIMIENTOS

Est.	Positivo	Neutral	Negativo	Ideas
mín	0.02	0.01	0.00	5
máx	0.99	0.50	0.95	14
$\bar{x}$	0.42	0.15	0.43	8.61
σ	0.25	0.09	0.22	2.54
P_{50}	0.37	0.13	0.43	9

En cuanto a la proporción de sentimientos, se observa que la mediana del sentimiento negativo es la más alta (0.43), seguida del sentimiento positivo (0.37) y, finalmente, el sentimiento neutral (0.13). Ademas, la variabilidad entre valores maximos y minimos de los tres sentimientos infican que existe cambios bruscos entre las ideas. Por ultimo la desviacion estandar indica que el sentimiento con menor variabilidad es neutral.

La Figura 5 muestra un ejemplo de cómo evolucionan los sentimientos a medida que se desarrolla cada idea dentro del discurso. Se evidencia la alta variabilidad de sentimientos postivo y negativos entre ideas y el sentimiento neutral mantiene una proporción relativamente baja a lo largo del discurso.

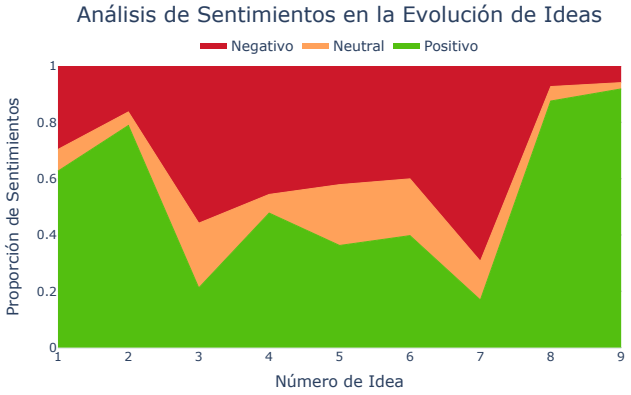

Figura 5. Análisis de Sentimientos en la Evolución de Ideas. La proporción de cada sentimiento se muestra a lo largo de las ideas sucesivas en el discurso.

Los resultados muestran que el predominio de sentimientos negativos en los discursos refleja un posible enfoque en temas sensibles, común en este contexto. La segmentación en ideas permite capturar la variabilidad emocional, facilitando un análisis que respete las complejidades emocionales e identificadon secciones criticas en el discurso.

C. *Validación de Hechos*

Tal como se menciona en la sección anterior, el objetivo es obtener resultados que vayan más allá de los aciertos y errores que el modelo pueda generar en sus predicciones, explorando también la capacidad del modelo para manejar declaraciones complejas y engañosas.

1). Matriz de Confusión: La Figura 6 muestra el desempeño del modelo en las tres categorías definidas en el conjunto de datos de entrenamiento: verdadero, falso e impreciso. Aunque el modelo presenta un número considerable de aciertos tanto en declaraciones verdaderas como en declaraciones falsas, se observan mayores errores en la categoría de declaraciones engañosas, donde el modelo tiende a clasificarlas principalmente como verdaderas.

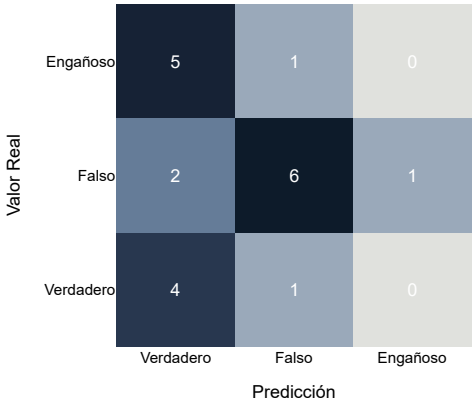

Figura 6. Matriz de Confusión para el módulo de validación de hechos.

Este comportamiento es, hasta cierto punto, esperable debido a la naturaleza de las declaraciones engañosas. Estas suelen incluir datos verídicos, pero combinados con falsedades o descontextualizaciones que no reflejan la actualidad. La mezcla de información verdadera y falsa en las declaraciones engañosas podría confundir al modelo, llevándolo a clasificarlas incorrectamente como verdaderas.

2). Similitud Coseno: Los resultados del cálculo de distancia coseno entre las validaciones y las respuestas generadas se presentan en la Figura 7. En esta figura se observa que las respuestas generadas para declaraciones falsas son las que más se asemejan al texto de las verificaciones reales, con una mediana de similitud de 0.824.

El análisis de similitud indica que el tono y estilo de las respuestas generadas son generalmente consistentes con las verificaciones reales. El texto de las afirmaciones tanto verdaderas como falsas muestran una alta similitud con el texto original, esto sugiere que el modelo es capaz de replicar el estilo y estructura de las respuestas de referencia de manera efectiva.

Sin embargo, las respuestas generadas en casos de error presentan una mayor variabilidad y una dispersión más amplia. Esto indica que el modelo tiene dificultades para mantener una consistencia en el tono y fuentes citadas ya que estaria alucinando en el texto generado.

3). Métricas de Calidad: Para el nivel de evaluación más alto, los resultados de las cuatro métricas propuestas se

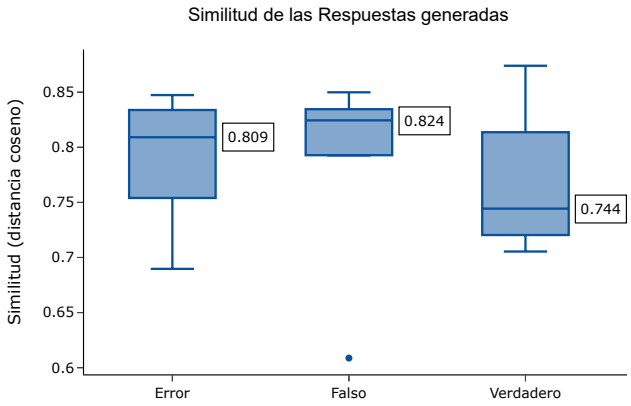

Figura 7. Similitud de las Respuestas Generadas en el módulo de validación de hechos, medida por la distancia coseno.

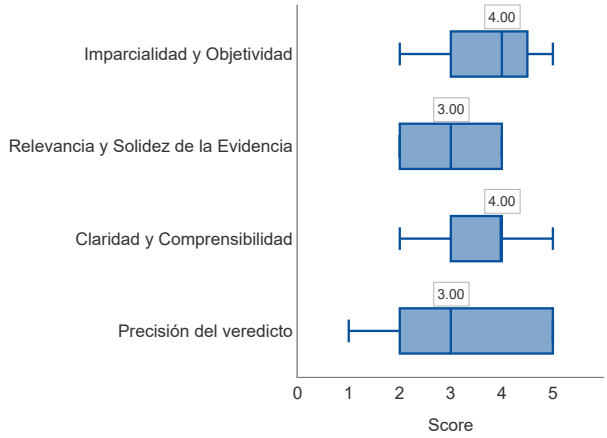

Figura 8. Evaluación de Métricas de Calidad.

resumen en la Figura 8. Como se mencionó en la sección anterior, la escala de evaluación va de 1 a 5.

En la dimensión de *precisión del veredicto* se observa una mediana de 3 con una alta dispersión. Este comportamiento se puede relacionar con lo visto en la matriz de confusión: las declaraciones de tipo engañoso son las que el modelo tiene más dificultad para clasificar correctamente, lo que hace que no alcance un nivel óptimo de precisión al generar su veredicto.

La métrica de *claridad y comprensibilidad* presenta una mediana de 4, lo que indica que, en la mayoría de los casos, el texto generado es fácil de comprender y sigue un orden estructurado. La baja dispersión sugiere que este comportamiento es constante en la mayoría de las pruebas realizadas.

En cuanto a *relevancia y solidez de la evidencia*, se obtiene una mediana de 3 con una alta dispersión o variabilidad en las pruebas realizadas. Este resultado indica que, aunque los datos y argumentos generados para justificar el veredicto son aceptables en general, en algunos casos pueden no ser tan relevantes o sólidos como en la validación real.

Por último, en la dimensión de *imparcialidad y objetividad* se observa una mediana de 4 con una baja dispersión, lo que indica que el texto generado mantiene un tono neutral y los sesgos son bajos en las pruebas realizadas, mostrando una consistencia en el desempeño de esta métrica.

V. Trabajos Futuros

La principal oportunidad de mejora en la validación de hechos, como se ha observado en las distintas evaluaciones, radica en optimizar la identificación de información engañosa. Además de incrementar el volumen de datos de entrenamiento, sería útil implementar técnicas adicionales, como el uso de agentes bajo el enfoque de ReAct, para mejorar la capacidad del modelo en el proceso de identificar información ambigua. Esto podría reducir la tendencia del LLM a clasificar información engañosa como verdadera y fortalecer su precisión.

Además, es muy importante encontrar un método para mejorar la explicabilidad del modelo. Si bien las respuestas incluyen la referencia de los datos empleados, sería idóneo contar con mecanismos que ayuden a entender cómo se ha llegado a las conclusiones del texto generado en la validación.

Este enfoque ha sido desarrollado en base a texto, pero actualmente los LLMs están evolucionando para ser multimodales. Por lo tanto, un siguiente paso sería encontrar un modelo adecuado para incorporar los audios de los discursos y, de esta manera, enriquecer el análisis del discurso.

VI. Conclusiones

Este trabajo presenta una aplicación práctica de Modelos de Lenguaje de Grande (LLMs) en el análisis de discursos políticos en Ecuador, abarcando tres tareas fundamentales: generación de resúmenes, análisis de sentimientos y verificación de hechos. Los resultados obtenidos destacan la utilidad de los módulos de resumen y análisis de sentimientos para hacer accesible información política compleja, permitiendo así una comprensión más clara de los discursos públicos.

El enfoque híbrido propuesto para el módulo de verificación de hechos ofrece resultados prometedores, siendo efectivo en la validación de declaraciones clasificadas como verdaderas y falsas en múltiples pruebas. Sin embargo, el modelo mostró limitaciones significativas en la detección de declaraciones engañosas, lo que indica una oportunidad de mejora en el tratamiento de afirmaciones ambiguas y contextualmente complejas.

A pesar de estas limitaciones, los resultados sugieren que los LLMs pueden contribuir a la democratización de la información y al fortalecimiento de la credibilidad en el discurso político, proporcionando herramientas que capaciten a los votantes para combatir la desinformación y

promover un debate público más informado. La constante evolución de los LLMs en los últimos años plantea nuevas posibilidades para abordar las limitaciones identificadas en este estudio mediante enfoques contextuales más avanzados, permitiendo así mejorar la precisión y neutralidad de los textos generados en aplicaciones futuras.

Referencias

[1] S. Banerjee, A. Agarwal, and S. Singla, "Llms will always hallucinate, and we need to live with this," 2024. [Online]. Available: https://arxiv.org/abs/2409.05746

[2] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," 2023. [Online]. Available: https://arxiv.org/abs/2210.03629

[3] A. Espinel, "Las fake news divulgadas en facebook y su incidencia en la opinión pública en las elecciones generales ecuador 2021," 2023.

[4] M. Linegar, R. Kocielnik, and R. M. Alvarez, "Large language models and political science," *Frontiers in Political Science*, vol. 5, 2023. [Online]. Available: https://www.frontiersin.org/journals/political-science/articles/10.3389/fpos.2023.1257092

[5] R. Patil and V. Gudivada, "A review of current trends, techniques, and challenges in large language models (llms)," *Applied Sciences*, vol. 14, no. 5, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/5/2074

[6] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "Gpt understands, too," 2023. [Online]. Available: https://arxiv.org/abs/2103.10385

[7] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, "Unleashing the potential of prompt engineering in large language models: a comprehensive review," 2024. [Online]. Available: https://arxiv.org/abs/2310.14735

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021. [Online]. Available: https://arxiv.org/abs/2005.11401

[9] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2312.10997

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[11] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

[12] L. Rettenberger, M. Reischl, and M. Schutera, "Assessing political bias in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2405.13041

[13] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto, "Whose opinions do language models reflect?" 2023. [Online]. Available: https://arxiv.org/abs/2303.17548

[14] C. Chen and K. Shu, "Combating misinformation in the age of llms: Opportunities and challenges," 2023. [Online]. Available: https://doi.org/10.1002/aaai.12188

[15] M. Li, B. Peng, M. Galley, J. Gao, and Z. Zhang, "Self-checker: Plug-and-play modules for fact-checking with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2305.14623

[16] X. Zhang and W. Gao, "Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method," 2023. [Online]. Available: https://arxiv.org/abs/2310.00305

[17] B. A. Galitsky, "Truth-o-meter: Collaborating with llm in fighting its hallucinations," *Preprints*, July 2023. [Online]. Available: https://doi.org/10.20944/preprints202307.1723.v1

[18] E. Hoes, S. Altay, and J. Bermeo, "Leveraging chatgpt for efficient fact-checking," *PsyArXiv. April*, vol. 3, 2023.

[19] O. Ovadia, M. Brief, M. Mishaeli, and O. Elisha, "Fine-tuning or retrieval? comparing knowledge injection in llms," 2024. [Online]. Available: https://arxiv.org/abs/2312.05934

[20] X. V. Lin, X. Chen, M. Chen, W. Shi, M. Lomeli, R. James, P. Rodriguez, J. Kahn, G. Szilvasy, M. Lewis, L. Zettlemoyer, and S. Yih, "Ra-dit: Retrieval-augmented dual instruction tuning," 2024. [Online]. Available: https://arxiv.org/abs/2310.01352

[21] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez, "Raft: Adapting language model to domain specific rag," 2024. [Online]. Available: https://arxiv.org/abs/2403.10131

[22] OpenAI, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[23] A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou, E. Wang, and X. Dong, "Better zero-shot reasoning with role-play prompting," 2024. [Online]. Available: https://arxiv.org/abs/2308.07702

[24] Lik Xun Yuan, "distilbert-base-multilingual-cased-sentiments-student (revision 2e33845)," 2023. [Online]. Available: https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student

[25] Ecuador Verifica, "¿quiénes somos?" 2024, accessed: 2024-08-15. [Online]. Available: https://ecuadorverifica.org/quienes-somos/

[26] Observatorio Legislativo, "Sobre nosotros - ecuador chequea," 2024, accessed: 2024-09-03. [Online]. Available: https://observatoriolegislativo.ec/sobre-nosotros/

[27] OpenAI, "Fine tuning - docs," 2024, accessed: 2024-11-03. [Online]. Available: https://platform.openai.com/docs/guides/fine-tuning

[28] A. Celikyilmaz, E. Clark, and J. Gao, "Evaluation of text generation: A survey," 2021. [Online]. Available: https://arxiv.org/abs/2006.14799

[29] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023. [Online]. Available: https://arxiv.org/abs/2306.05685

[30] A. Anthropic, "Claude 3.5 sonnet model card addendum," *Claude-3.5 Model Card*, 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-5-sonnet

[31] B. Lechon, "Analizador de discursos," GitHub repository, 2024. [Online]. Available: https://github.com/balechon/AnalizadorDiscursos