UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Smart Citizen Control of Public Procurement in Ecuador: Classification of accusatory comments from "Sistema Oficial de Contratación Pública del Ecuador (SOCE)" using LLMs

Proyecto de Titulación

Bryan Patricio Núñez Alverca

Felipe Grijalva, 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, 02 de diciembre de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ
COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Smart Citizen Control of Public Procurement in Ecuador: Classification of accusatory comments from "Sistema Oficial de Contratación Pública del Ecuador (SOCE)" using LLMs

Bryan Patricio Núñez Alverca

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, diciembre 2024

© DERECHOS DE AUTOR

Nombre del estudiante: Bryan Patricio Núñez Alverca

Código de estudiante: 00339646

C.I.: 1804908646

Lugar y fecha: Quito, 02 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UNPUBLISHED DOCUMENT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

Dedico este esfuerzo a mi sobrino, Martin Patricio tercero, con la esperanza de que encuentre en estas páginas una chispa que le inspire a mirar el mundo con otros ojos, esos que ven lo extraordinario en lo cotidiano y descubren la magia en los detalles más pequeños. Que nunca pierda la curiosidad por todo lo que le rodea, desde las maravillas más diminutas hasta los misterios más grandes del universo. Deseo que siempre tenga el valor de hacer preguntas, de explorar sin límites y de buscar la libertad en todas sus formas: la libertad de pensar, de soñar y de elegir su propio camino. Que su espíritu inquieto nunca se detenga y que cada paso que de esté guiado por la pasión y el deseo de comprender el mundo y transformarlo con sus propias manos. Él es una fuente muy grande de inspiración y orgullo, y espero que esta dedicación sea un recordatorio de que en cada rincón del conocimiento hay un mundo esperando ser descubierto, por ti y para ti. Con cariño,
Tu tío.

AGRADECIMIENTOS

A mis padres, Verónica y Patricio, quienes con su amor incondicional, esfuerzo y apoyo han sido el cimiento de todo lo que he logrado. Gracias por enseñarme el valor del trabajo, la perseverancia y, sobre todo, por creer en mí incluso en los momentos más desafiantes. A mi hermano Diego, quien siempre ha sido un ejemplo de fortaleza y determinación. Gracias por tus palabras de aliento, por estar presente en cada etapa de este camino y por ser una fuente constante de inspiración. Tu apoyo incondicional me ha dado la fuerza para superar cualquier obstáculo, y por ello, siempre te estaré profundamente agradecido. A mi hermana Karla, quien desde Brasil siempre ha estado presente con su cariño y palabras de aliento, a quien envío un agradecimiento especial por demostrarme que las distancias nunca separan los lazos que realmente importan. A mi cuñada, María José, por su amabilidad y apoyo, y a mis sobrinos Yared y Martín, quienes son una gran fuente de inspiración y alegría, recordándome siempre la importancia de mirar al futuro con esperanza. Finalmente, a Jazmín, mi compañera, quien ha caminado a mi lado en este viaje, brindándome su amor, comprensión y apoyo en cada paso del camino. Tu presencia ha hecho de este proceso algo especial, y por ello, te estaré siempre agradecido. A todos ustedes, gracias por ser mi fuerza, mi inspiración y mi motivo para seguir adelante.

RESUMEN

En el contexto de las compras públicas en Ecuador, donde la transparencia y la equidad son fundamentales, este estudio aborda el desafío de identificar comentarios acusatorios realizados por los proveedores durante la etapa de preguntas y respuestas en los procesos de licitación. Se recopilaron un total de 5,005 frases del portal de compras públicas, de las cuales solo 147 fueron etiquetadas como acusatorias, evidenciando un desbalance significativo entre clases. Se evaluaron cuatro modelos de lenguaje de gran escala (LLMs): ChatGPT 4o-mini, Llama 3.1 (8B y 70B) y PHI 3.5 mini, utilizando técnicas avanzadas de prompting como greedy search y beam search. Los resultados muestran que ChatGPT 4o-mini logró el mejor desempeño, con un AUC de 0.97222 y un recall del 96.55% para la clase acusatoria, seguido de cerca por Llama 3.1 8B con un AUC de 0.96099 y un recall del 93.10%. Aunque Llama 3.1 70B presentó un rendimiento competitivo con un AUC de 0.94611 y un recall del 90.00%, sus resultados se vieron influenciados por el tamaño del modelo, lo que requirió el uso de técnicas de optimización de recursos, como la cuantización a 4 bits y la reducción de la ventana de contexto. PHI 3.5 mini mostró un rendimiento significativamente inferior, con un recall del 75.86%. A pesar del liderazgo de ChatGPT, se recomienda Llama 3.1 8B debido a su desempeño comparable y su naturaleza de código abierto, que permite su implementación en infraestructura local sin los costos asociados al uso de API. Este estudio destaca la importancia de los modelos de lenguaje en la detección temprana de posibles irregularidades en las compras públicas, ofreciendo un enfoque escalable y eficiente para mejorar la transparencia y la supervisión gubernamental en Ecuador.

Palabras clave: LLM, búsqueda codiciosa, búsqueda en haces, Chat-GPT, Llama, PHI.

ABSTRACT

In the context of public procurement in Ecuador, where transparency and fairness are essential, this study addresses the challenge of identifying accusatory comments made by suppliers during the question-and-answer stage of bidding processes. A total of 5,005 phrases were collected from the public procurement portal, of which only 147 were labeled as accusatory, highlighting a significant class imbalance. Four large language models (LLMs) were evaluated: ChatGPT 4o-mini, Llama 3.1 (8B and 70B), and PHI 3.5 mini, leveraging advanced prompting techniques such as greedy search and beam search. The results show that ChatGPT 4o-mini achieved the best performance, with an AUC of 0.97222 and a recall of 96.55% for the accusatory class, followed closely by Llama 3.1 8B with an AUC of 0.96099 and a recall of 93.10%. Although Llama 3.1 70B demonstrated competitive performance with an AUC of 0.94611 and a recall of 90.00%, its results were influenced by the model's size, necessitating the use of resource optimization techniques such as 4-bit quantization and a reduction in the context window. PHI 3.5 mini showed significantly lower performance, with a recall of 75.86%. Despite ChatGPT's leadership, Llama 3.1 8B is recommended due to its comparable performance and open-source nature, which allows implementation on local infrastructure without the costs associated with API usage. This study highlights the importance of language models in the early detection of potential irregularities in public procurement, offering a scalable and efficient approach to enhance transparency and government oversight in Ecuador.

Key words: LLM, greedy search, beam search, Chat-GPT, Llama, PHI.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Smart Citizen Control of Public Procurement in Ecuador: Classification of accusatory comments from "Sistema Oficial de Contratación Pública del Ecuador (SOCE)" using LLMs

Felipe Grijalva, *Senior Member, IEEE,* Bryan Núñez, *Member, IEEE,*

***Abstract*—In the context of public procurement in Ecuador, where transparency and fairness are essential, this study addresses the challenge of identifying accusatory comments made by suppliers during the question-and-answer stage of bidding processes. A total of 5,005 phrases were collected from the public procurement portal, of which only 147 were labeled as accusatory, highlighting a significant class imbalance. Four large language models (LLMs) were evaluated: ChatGPT 4o-mini, Llama 3.1 (8B and 70B), and PHI 3.5 mini, leveraging advanced prompting techniques such as greedy search and beam search. The results show that ChatGPT 4o-mini achieved the best performance, with an AUC of 0.97222 and a recall of 96.55% for the accusatory class, followed closely by Llama 3.1 8B with an AUC of 0.96099 and a recall of 93.10%. Although Llama 3.1 70B demonstrated competitive performance with an AUC of 0.94611 and a recall of 90.00%, its results were influenced by the model's size, necessitating the use of resource optimization techniques such as 4-bit quantization and a reduction in the context window. PHI 3.5 mini showed significantly lower performance, with a recall of 75.86%. Despite ChatGPT's leadership, Llama 3.1 8B is recommended due to its comparable performance and open-source nature, which allows implementation on local infrastructure without the costs associated with API usage. This study highlights the importance of language models in the early detection of potential irregularities in public procurement, offering a scalable and efficient approach to enhance transparency and government oversight in Ecuador.**

***Index Terms*—LLM, greedy search, beam search, ChatGPT, Llama, PHI**

I. Introduction

IN Ecuador, the context of public procurement is a critical area that demands high levels of transparency and fairness to ensure that contracting processes are just and accessible to all interested suppliers. However, the country faces significant challenges in this area, as evidenced by its position in the 2023 Corruption Perceptions Index, where it ranked 115th globally [1]. Limitations in transparency mechanisms have allowed corrupt activities and unfair practices to go unnoticed, undermining trust in public institutions [2]. This issue is particularly concerning during the question-and-answer phase in public tenders, where suppliers, rather than merely asking questions, may express concerns or suspicions of favoritism, bias, and other irregularities in the bidding process [3].

In collaboration with the Universidad San Francisco de Quito (USFQ) and the German Development Cooperation (GIZ), Kapak was developed—a web portal designed to analyze public procurement data and detect potential corruption risks [4]. However, Kapak has certain limitations, as its corruption risk indicators are only calculated after the information has been captured and processed, often resulting in delayed interventions by authorities. This project aims to enhance early detection capabilities by integrating advanced language models for comment analysis, which could reveal potential signs of corruption before they escalate into major issues.

In this context, large language models (LLMs) such as ChatGPT 4o-mini, Llama 3.1 (8B and 70B), and PHI 3.5 mini have become powerful tools for text analysis and classification. To optimize generation and classification, techniques like greedy search and beam search are utilized, enabling these models to select the best word sequences based on their probability. These decoding techniques have proven effective in enhancing coherence and accuracy in response generation, maximizing efficiency in tasks requiring natural language analysis [5], [6]. Additionally, 4-bit quantization in models like Llama 3.1 70B has facilitated the reduction of computational resource

F. Grijalva and B. Núñez are with Universidad San Francisco de Quito USFQ

usage, allowing these models to be employed in hardware-constrained environments without significantly compromising performance [7].

The challenge lies in the accurate and efficient classification of statements made by suppliers on the public procurement portal, where some questions may contain direct or indirect accusations regarding the impartiality of the process. To address this issue, 5,005 statements were collected and labeled as accusatory or non-accusatory. These statements capture not only the literal content but also the nuances of Ecuadorian language, including sarcasm, double meanings, and other expressions that may reflect potential accusations of bias or favoritism.

The primary objective of this project is to develop a robust classifier model capable of automatically and accurately identifying accusatory comments in real time. This capability will enable early detection of potential signs of corruption, facilitating quicker interventions and improving transparency in public procurement processes in Ecuador. By optimizing the automated analysis of these comments, the project aims not only to increase the efficiency of government oversight but also to strengthen the trust of suppliers and the public in the integrity of public contracting processes.

II. Prior Works

The project presented by Torres Berrú [8] employs data mining and natural language processing techniques to identify patterns of overpricing and favoritism in public contracts in Ecuador. This approach is primarily based on the quantitative analysis of structured data to detect irregularities in contracting processes. In contrast, my research utilizes large language models (LLMs) such as ChatGPT and Llama, leveraging advanced prompting techniques (like greedy search and beam search) to analyze textual comments made by suppliers during the question-and-answer stage. Unlike Torres Berrú's numerical approach, my project focuses on interpreting qualitative content, capturing linguistic nuances such as sarcasm and indirect accusations that may go unnoticed in traditional analyses based solely on quantitative data.

The study "Automatic Procurement Fraud Detection with Machine Learning" [9] applies neural networks to detect fraud in procurement processes by using quantitative features of transactions within a specific business environment (SF Express). While this approach relies on binary classification algorithms to identify suspicious patterns, my project goes further by focusing on the classification of textual comments in the context of public procurement in Ecuador. Here, the analysis is not limited to predefined data but leverages prompting techniques in LLMs to interpret the implicit meanings within suppliers' comments, enabling the identification of potential signs of corruption in real time through a more contextual and linguistic approach.

Finally, the article "A Machine Learning Model to Identify Corruption in Mexico's Public Procurement Contracts" [10] uses classifiers such as random forests to predict the likelihood of corruption by analyzing the relationship between buyers and suppliers based on structured data. In contrast, my project focuses on a more detailed qualitative analysis using LLMs to assess the accusatory nature of comments on a scale from 1 to 10. This approach enables a more precise evaluation of language subtleties, such as insinuations or veiled accusations, which cannot be captured by models exclusively focused on numerical and structured variables. Thus, my research complements these traditional approaches by integrating a deep natural language analysis to detect potential acts of corruption at earlier stages.

III. Materials and Methods

In this section, we will first provide a detailed description of the dataset provided by the Kapak platform, which includes comments collected from Ecuador's public procurement portal, along with the labeling process [4]. Next, we will explain the stages of our proposed methodology for classifying accusatory and non-accusatory comments using large language models (LLMs). Finally, we will present the experimental setup, including the parameters and techniques used, to evaluate the performance of our models in the classification task.

A. *Dataset*

The dataset used in this study was provided by the Kapak platform, which collects comments and questions from Ecuador's public procurement portal. The goal of this dataset is to capture the concerns, doubts, and potential accusations of suppliers towards government entities during the public bidding process. Given the importance of transparency in public contracting, these comments serve as a valuable source for identifying potential irregularities or signs of favoritism.

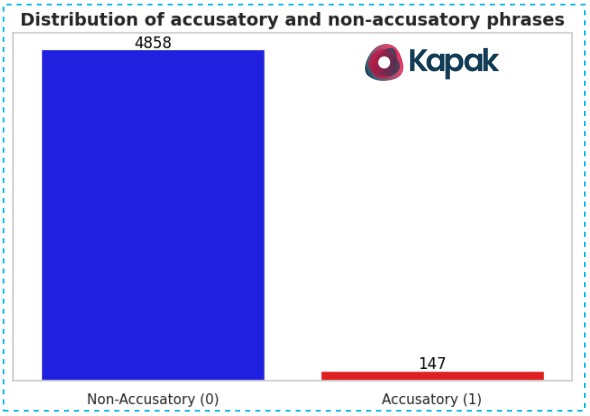

Figure 1. Distribution of accusatory and non-accusatory phrases in the dataset

The dataset includes a total of 5,005 comments, of which 4,858 have been classified as non-accusatory, while 147 have been labeled as accusatory. This distribution reveals a clear imbalance, where only a small percentage of comments contain direct or indirect accusations. The labeling was conducted using criteria defined by experts on the Kapak platform, based on the content and context of each comment.

Figure 1 illustrates this distribution, highlighting the significant difference between the number of non-accusatory and accusatory comments. This imbalance presents a challenge for the classification model, as it requires advanced techniques to handle the disparity and improve accuracy in detecting accusatory phrases which, although less frequent, are crucial for identifying potential corruption risks.

B. *Proposed method*

Our approach consists of several stages, as shown in figure 2. We utilized a dataset provided by Kapak, containing comments from Ecuador's public procurement portal, categorized as accusatory and non-accusatory. Given the class imbalance, we applied a data augmentation process with ChatGPT to generate additional accusatory comments and balance the dataset [11].

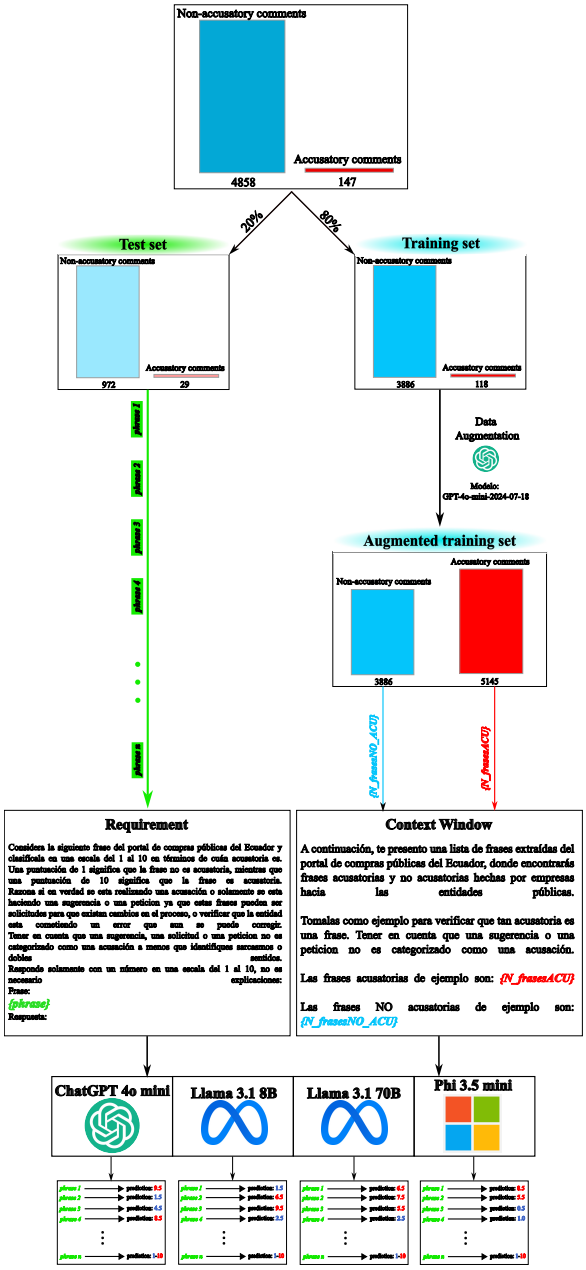

Figure 2. Workflow of data processing, augmentation, and model evaluation

After splitting the dataset into training and testing sets, we constructed a context window using labeled examples to guide the models in classifying comments on a scale from 1 to 10 based on their level of accusatoriness. Instead of adjusting model weights, prompting techniques were applied using strategies such as greedy search and beam

search [12] to optimize the responses generated by ChatGPT 4o-mini, Llama 3.1 (8B and 70B), and PHI 3.5 mini models.

The following sections detail how these techniques were employed in the classification process.

C. *Dataset Split*

The original dataset provided by the Kapak platform was segmented into 80% for training and 20% for testing using a stratified split. This ensured that both subsets maintained the original proportion of accusatory and non-accusatory comments.

D. *Data Augmentation Using Prompting Techniques*

Given the significant class imbalance with a predominance of non-accusatory comments over a minority of accusatory ones advanced prompting techniques were employed to augment the volume of data for the minority class. Instead of using the entire dataset, the approach focused exclusively on accusatory phrases from the training set, instructing the model to generate similar accusatory statements based on the provided examples.

- **Model**: ChatGPT 4o-mini
- **Context window**: A specific context was created based on accusatory phrases extracted from the training dataset.
- **Parameters**:
  - **Temperature**: 0.5, to balance creativity and coherence.
  - **Maximum tokens**: 50, to ensure the phrases were informative without being overly lengthy.
  - **Seed**: 42, to guarantee reproducibility.
  - **Parallelization**: An additional 5,027 phrases were generated using concurrent threads to optimize execution time.

As shown in Figure 3, the context window includes the variable `frasesACU`, which contains all the accusatory phrases from the training set.

Context Window for Data Augmentation

En el contexto de compras públicas del Ecuador, existe un proceso de preguntas que realizan las empresas que desean participar en los concursos. Estas preguntas están dirigidas a las entidades públicas que están creando el concurso o la necesidad. En esta etapa de preguntas y respuestas, las empresas que desean participar en el proceso realizan comentarios respecto a dudas en el proceso. Sin embargo, en algunos casos, dichas empresas notan que los procesos están siendo direccionados para que una empresa en particular gane el proceso, o también, para que solo un tipo de marca o producto en específico pueda cumplir con las especificaciones de la necesidad, lo que causa que el acceso a la participación no sea totalmente libre, lo cual es ilegal en este tipo de compras en el estado ecuatoriano.

A continuación, te presento una lista de frases extraídas del portal de compras públicas del Ecuador, donde encontrarás frases acusatorias hechas por empresas hacia las entidades públicas. Tu tarea es generar nuevas frases acusatorias similares a las que te presento, tomando en cuenta las siguientes características:
- Las nuevas frases deben ser originales y no copias directas de las frases de ejemplo ni de las frases generadas anteriormente.
- Las frases deben parecer escritas por una persona real y reflejar preocupaciones específicas sobre la imparcialidad del proceso.
- Las frases acusatorias pueden enfocarse en la especificidad del personal solicitado, la exclusividad de marcas o productos, o cualquier indicio de que el proceso está dirigido a favorecer a un competidor particular.
- Usa un tono formal y profesional apropiado para el contexto de compras públicas.
- No repitas los mismos fabricantes en múltiples frases para asegurar diversidad.

Las frases de ejemplo son: "*{frasesACU}*

Figure 3. Context window for data augmentation

E. *Roles in the Prompting Strategy: Context Window and Requirement*

- **Role "system" (Context Window)**:
  - Establishes the general context that the model must consider before performing the classification. This includes detailed examples of both accusatory and non-accusatory phrases, helping the model understand the tone and specific nuances of the language used in the public procurement portal.
  - The context window is crucial for the model to capture subtle details such as sarcasm and indirect accusations.
- **Role "user" (Requirement)**:
  - Provides a specific instruction for the classification task. Here, the model must classify a phrase on a scale from 1 to 10, indicating how accusatory it is. The requirement includes guidelines on how to distinguish between an accusation and a mere suggestion or request.
  - This ensures that the model not only grasps the background context but also performs the task accurately.

*1) **Decoding Methods in Language Models**:* Two decoding techniques are described here: greedy search and beam search, which were the methods explored in this project.

- **Greedy Search**
  Greedy search is a simple and fast decoding technique that, at each step, selects the word with the highest probability as the next word in the generated sequence [13]. This means the model deterministically chooses the most likely token at each step, without considering alternatives in future selections.
  - It is quick and efficient in terms of execu-

tion time, as it only needs to consider one option per step.
  - It only takes into account the probability of the next token without evaluating alternative sequences, which may result in outputs that are not globally optimal.
  - Depending on the model, it is not ideal when text generation needs to meet specific constraints or conditions.

- **Beam Search and Forced Words**
  Beam search is a more advanced decoding technique that allows the model to explore multiple sequences in parallel before making a final decision [14]. Unlike greedy search, beam search maintains a fixed number of options (referred to as "beams") at each step, evaluating the most promising sequences and selecting the one with the highest cumulative probability at the end.
  - It enables the exploration of several possible sequences simultaneously, reducing the likelihood of the model getting stuck in a suboptimal choice.
  - It is particularly useful when outputs need to meet specific constraints, such as, in our case, ensuring the model returns a number between 1 and 10.
  - A key feature of beam search is its ability to enforce the inclusion of specific words or tokens in the generated output. This is achieved using the "forced words" option, which, in models implemented with Hugging Face, is only supported when using beam search.
  - In this project, beam search with forced words was used to ensure that the PHI 3.5 mini model consistently returned a number between 1 and 10 as the classification result. This was crucial because, when using greedy search, the model occasionally failed to generate a numerical response.

*2) **Models and Parameter Optimization**:* During the evaluation process of the models ChatGPT 4o-mini, Llama 3.1 8B, Llama 3.1 70B, and PHI 3.5 mini, various configurations of context windows were tested, adjusting the number of accusatory and non-accusatory phrases to maximize accuracy without exceeding the available resources.

For the Llama and PHI models, servers equipped with NVIDIA A100 GPUs were used, which allowed for handling larger context windows and maximizing token processing capacity. However, due to its considerable size, Llama 3.1 70B had to be quantized to 4 bits to reduce memory and resource usage. Even so, this model faced GPU memory limitations, which necessitated reducing the context window size compared to Llama 3.1 8B and PHI 3.5 mini, which could handle larger context windows without issues.

In contrast, ChatGPT 4o-mini was evaluated at the end of the process, as its use involves a cost per token through the OpenAI API. For this reason, a baseline was first established using the Llama 3.1 8B models to optimize the results before incurring additional costs. Additionally, since ChatGPT was accessed via the API and not loaded onto a local GPU, it did not face memory constraints, allowing for the use of larger context windows adjusted to the parameters tested on the Llama models.

Llama 3.1 8B

Llama 3.1 8B is part of the Llama model series developed by Meta AI. It is built on the standard Transformer architecture [15], which enables it to perform advanced text comprehension and generation tasks across multiple languages and domains. Unlike previous versions, the Llama 3.1 series introduces improvements in data quality and diversity, as well as enhanced training scale. This includes the capability to handle extended context windows and better support for tasks like reasoning and coding [16].

The Llama 3.1 8B model is an optimized variant for instruction-based tasks, trained on a high-quality dataset spanning various domains, allowing it to deliver outstanding performance on benchmarks such as MMLU and HumanEval. Its architecture is based on a dense model with 32 layers, a dimension of 4,096, and 32 attention heads, providing a balance between efficiency and generalization capabilities [16].

3) Optimal Parameter Configuration for best results: The Llama 3.1 8B model was loaded and executed on an NVIDIA A100 GPU with support for `bfloat16` (16-bit floating-point tensors), optimizing both inference speed and efficient memory usage. During evaluation, the following configurations were used to achieve optimal results:

- **Number of phrases in the context window**:
  - Accusatory phrases: 250
  - Non-accusatory phrases: 250
- **Temperature**: 0.5
- **Top-p**: 0.5
- **Maximum number of generated tokens** (`max_new_tokens`): 10

Llama 3.1 70B

Llama 3.1 70B is part of the Llama model series developed by Meta AI. Like its smaller counterparts, it is based on the standard Transformer architecture [16], optimized for advanced text generation and comprehension tasks across various languages and contexts. However, this version takes the series' capabilities to a higher level, with 70 billion parameters, enabling much deeper and more detailed text analysis [16].

This model is designed to handle complex tasks that require a high degree of precision and reasoning, achieving outstanding performance on multiple benchmarks such as MMLU and HumanEval. The architecture of Llama 3.1 70B consists of 80 layers, a dimension of 8,192, and 64 attention heads, which provide exceptional capacity to process contextual information and generate detailed responses [16].

4) Optimal Parameter Configuration for best results:

- **Quantization**: 4-bit (`load_in_4bit=True`) to optimize GPU memory usage.
- **Number of phrases in the context window**:
  - Accusatory phrases: 100
  - Non-accusatory phrases: 100
- **Temperature**: 0.5 to balance diversity and coherence.
- **Top-p**: 0.5 to limit cumulative probability and promote diversity in responses.
- **Maximum number of generated tokens** (`max_new_tokens`): 10 to ensure concise responses.

PHI 3.5 mini

PHI 3.5 Mini is part of the language model series developed by Microsoft, designed to achieve performance comparable to much larger models, such as GPT-3.5 and Mistral 8x7B, but with a significantly smaller size. Based on the Transformer architecture, PHI 3.5 Mini has approximately 3.8 billion parameters and is optimized for text generation and comprehension tasks across multiple domains [17].

This model uses a hybrid data approach, combining high-quality filtered web data with synthetic data generated by language models, allowing it to reach competitive performance levels with larger models. Additionally, PHI 3.5 Mini has been fine-tuned for chat, reasoning, and other advanced applications using supervised fine-tuning and Direct Preference Optimization (DPO) techniques to enhance its alignment and safety in chat environments [17].

5) Optimal Parameter Configuration for best results: During the evaluation process, it was found that traditional generation using greedy search did not yield the desired results, as the model tended to generate explanations rather than a number between 1 and 10, which was required for the classification of accusatory phrases. To address this issue, beam search with forced words was employed to ensure that responses were exclusively numbers on the requested scale.

The configuration that provided the best results was as follows:

- **Number of phrases in the context window**:
  - Accusatory phrases: 250
  - Non-accusatory phrases: 250
- **Decoding strategy**: Beam search with 5 beams (`num_beams=5`)
- **Forced words**: Specific tokens (numbers 1 to 10) enforced using `force_words_ids`
- **Maximum number of generated tokens**: 5

ChatGPT 4o-mini

The ChatGPT 4o-mini model, developed by OpenAI, is an optimized variant that is accessed via the API rather than being loaded locally on a GPU [18]. Unlike the Llama and PHI models, which were run in an environment with an A100 GPU, this model uses remote servers provided by OpenAI to handle both the computational load and text generation. This results in a different approach to infrastructure optimization, focusing

on token limits and API constraints rather than local hardware.

6) Optimal Parameter Configuration: During the evaluation phase, the parameters for the ChatGPT model were adjusted to align with those used for Llama 3.1 8B, in order to ensure a fair comparison between both models. This was done after establishing a solid baseline with Llama, as the ChatGPT model was tested last due to the cost associated with tokens generated via the API.

The optimal configuration was as follows:

- **Number of phrases in the context window**:
  - Accusatory phrases: 250
  - Non-accusatory phrases: 250
- **Temperature**: 0.5
- **Top-p**: 0.5
- **Maximum number of generated tokens** (`max_new_tokens`): 10

IV. RESULTS AND DISCUSSION

The following section presents the results obtained in the classification of accusatory phrases within the context of public procurement, evaluating the performance of the models.

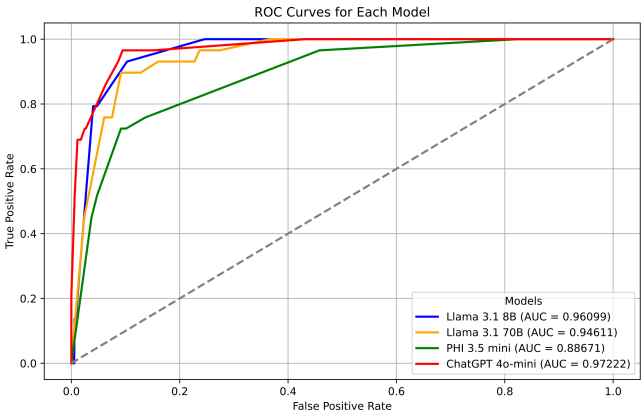

Figure 4. ROC curves for each model

The presented ROC curves (figure 4) highlight key differences in the discriminative capabilities of the evaluated models. ChatGPT 4o-mini, with an AUC of 0.97222, demonstrates the best performance, reflecting its ability to maintain a consistent balance between true positive and false positive rates. Its near-ideal curve indicates high recall and specificity, essential qualities for tasks where precise identification of positive class (accusatory phrases) is critical. The Llama 3.1 8B model follows with an AUC of 0.96099, showcasing its strong discriminative power. This model leverages its balanced architecture to achieve performance almost on par with ChatGPT. Llama 3.1 70B, with an AUC of 0.94611, shows a decline in discriminative capability, which can be attributed to context window constraints and the need for 4-bit quantization to manage its higher computational requirements. Finally, the PHI 3.5 mini model, with an AUC of 0.88671, exhibits significantly lower capacity to distinguish between the evaluated classes. Despite the use of advanced techniques such as beam search and forced words, its performance suggests reduced recall in identifying positive cases.

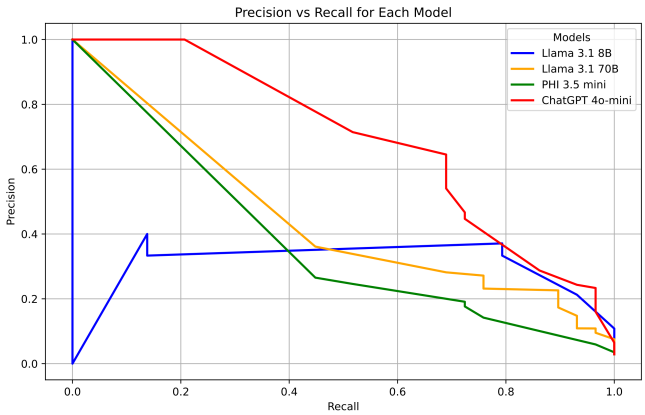

Figure 5. Presicion vs recall for each model

The Precision vs. Recall curve (figure 5) illustrates significant differences in the performance of the models. ChatGPT 4o-mini stands out with a gradual decline in precision as recall increases, maintaining stable performance at both extremes. This indicates its ability to handle both metrics in a balanced manner. On the other hand, Llama 3.1 8B shows an irregular curve at low recall values, reflecting potential inconsistencies in its performance, although it stabilizes at higher recall values. Llama 3.1 70B demonstrates a uniform decline in precision as recall increases. Lastly, PHI 3.5 mini exhibits a more pronounced drop in precision as recall is prioritized, indicating a lower capacity to maintain balance between the two metrics, especially at high recall values.

ChatGPT 4o-mini achieves a remarkable recall of 0.97 for the "Accusatory" class (table I), positioning itself as the most effective model for

Table I
RECALL VALUES FOR THE POSITIVE CLASS (ACCUSATORY) ACROSS ALL MODELS

Model	Recall
ChatGPT 4o-mini	0.97
Llama 3.1 8B	0.93
Llama 3.1 70B	0.90
PHI 3.5 mini	0.76

capturing phrases containing accusations. This level of recall is particularly critical in contexts where minimizing false negatives is essential to ensure that all potential accusatory phrases are identified. Similarly, Llama 3.1 8B achieves a recall of 0.93 for this class (table I), also demonstrating robust performance, albeit slightly lower than ChatGPT 4o-mini. In contrast, Llama 3.1 70B, despite its larger architecture, presents a recall of 0.90 for the "Accusatory" class (table I). This suggests that computational constraints and the need to reduce its context window impacted its ability to effectively discriminate this positive class. PHI 3.5 mini, with a recall of 0.76 for the "Accusatory" class (table I), exhibits limited performance compared to the other models, despite employing techniques like beam search and forced words to enhance its detection capabilities.

The confusion matrices reveal that the models exhibit variable performance in correctly identifying positive class (Accusatory), which is directly tied to the reported recall values. For ChatGPT 4o-mini, the model achieves a recall of 96.55% (figure 6a) for the Accusatory class, correctly identifying 28 true positives with only 1 false negative. This excellent recall performance indicates the model's strong ability to detect the majority of positive cases. However, it also records 92 false positives. The Llama 3.1 8B model achieves a recall of 93.10% (figure 6b) for the Accusatory class, with 27 true positives and 2 false negatives. This reflects a slight reduction in its ability to identify accusatory cases compared to ChatGPT 4o-mini. Llama 3.1 70B, on the other hand, shows a recall of 89.66% (figure 6c), correctly identifying 26 true positives but with 3 false negatives. Finally, PHI 3.5 mini exhibits the lowest performance, with a recall of 75.86% (figure 6d), correctly identifying only 22 true positives while accumulating 7 false negatives.

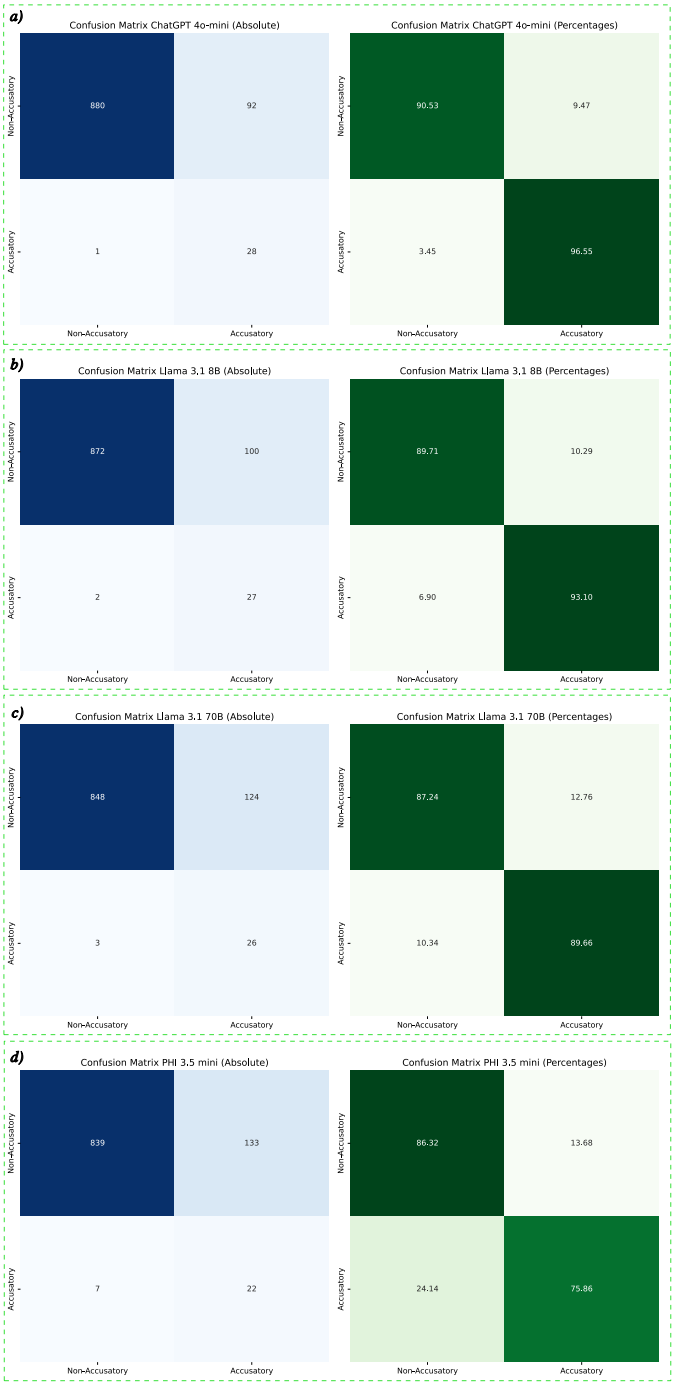

Figure 6. Confusion Matrices (Absolute and Percentage) for Model Performance: a) ChatGPT 4o-mini, b) Llama 3.1 8B, c) Llama 3.1 70B, d) PHI 3.5 mini

V. CONCLUSION

- The results demonstrate that ChatGPT 4o-mini is the model with the best overall performance for identifying accusatory phrases in the context of public procurement. With a recall of 96.55% and consistent precision across different metrics, it showcases its ability to minimize false negatives, which is critical

for ensuring early detection of potential corruption risks. This outcome highlights its suitability for tasks where sensitivity to positive cases is a priority.

- Llama 3.1 8B achieves a competitive recall of 93.10%, reflecting a good balance between computational efficiency and performance. However, Llama 3.1 70B, despite its more complex architecture, shows a negative impact on performance due to the need for quantization and reduced context window size, achieving a recall of 89.66%. On the other hand, PHI 3.5 mini, with a recall of 75.86%, demonstrates greater challenges in detecting positive cases, even with advanced techniques like beam search and forced words.
- The implementation of advanced techniques, such as data augmentation through prompting, the use of specific context windows, and parameter tuning for each model, played a crucial role in the observed performance. The optimization of configurations helped overcome inherent limitations in the models, such as class imbalance in the dataset and the need to adapt architectures to restricted computational resources.
- The recall metric for the positive class emerges as a critical factor in this study due to its direct relationship with identifying accusatory phrases. High recall values in models like ChatGPT 4o-mini and Llama 3.1 8B indicate their ability to detect a significant proportion of accusatory cases, reducing the likelihood of false negatives that could go unnoticed in real-world government oversight contexts. This underscores the importance of prioritizing this metric in tasks where missing a positive case could have significant consequences.
- While ChatGPT 4o-mini achieves the best overall performance, the use of Llama 3.1 8B is recommended due to its comparable performance, with a recall of 93.10% for the accusatory class, and the advantage of being an open-source model. This enables its implementation on local infrastructure, eliminating the costs associated with using the ChatGPT API and providing greater control over data and the inference process. These characteristics make it an efficient and accessible solution for contexts with budgetary and infrastructure constraints.

Acknowledgment

The authors express their sincerest gratitude to Universidad San Francisco de Quito (USFQ) for its invaluable academic and logistical support during the development of this project. Additionally, special recognition is extended to the KAPAK platform for providing the dataset necessary for this study and for its commitment to enhancing transparency in public procurement processes in Ecuador. This work would not have been possible without their collaboration and continuous support.

References

[1] "2023 corruption perceptions index: Explore the... - transparency.org," https://www.transparency.org/en/cpi/2023/index/ecu, (Accessed on 11/11/2024).

[2] M. R. A. Martinez, B. G. P. Andrade, T. M. L. Medina, and I. P. A. Bonilla, "Correlación entre el indice de percepción de la corrupción y su impacto en la inversión extranjera directa (ied) del año 2010 al 2022 en 164, 186 y 169 países," *Ciencia Latina Revista Científica Multidisciplinar*, vol. 7, no. 4, pp. 3287–3302, 2023.

[3] C. B. Ocampo León, "Análisis jurídico del problema del sistema nacional de contratación pública sobre la contraposición entre la ley orgánica del sistema nacional de contratación pública (losncp) el reglamento general del sistema nacional de contratación pública (rglosncp), el código organico administrativo (coa), y las resoluciones emitidas por el servicio nacional de contratación pública (sercop)," 2020.

[4] "Kapak - transparencia en compras públicas," https://kapak.usfq.edu.ec/#/inicio, (Accessed on 11/11/2024).

[5] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beam search: Decoding diverse solutions from neural sequence models," *arXiv preprint arXiv:1610.02424*, 2016.

[6] R. Leblond, J.-B. Alayrac, L. Sifre, M. Pislar, J.-B. Lespiau, I. Antonoglou, K. Simonyan, and O. Vinyals, "Machine translation decoding beyond beam search," *arXiv preprint arXiv:2104.05336*, 2021.

[7] Y. Zhao, C.-Y. Lin, K. Zhu, Z. Ye, L. Chen, S. Zheng, L. Ceze, A. Krishnamurthy, T. Chen, and B. Kasikci, "Atom: Low-bit quantization for efficient and accurate llm serving," *Proceedings of Machine Learning and Systems*, vol. 6, pp. 196–209, 2024.

[8] Y. Torres Berrú *et al.*, "Minería de datos y texto para la detección de fraude en contratos públicos: caso de estudio sistema oficial de contratación pública de ecuador," 2023.

[9] J. Bai and T. Qiu, "Automatic procurement fraud detection with machine learning," *arXiv preprint arXiv:2304.10105*, 2023.

[10] A. Aldana, A. Falcón-Cortés, and H. Larralde, "A machine learning model to identify corruption in m\'exico's public procurement contracts," *arXiv preprint arXiv:2211.01478*, 2022.

[11] Y. Wang, L. Guo, E. Shi, W. Chen, J. Chen, W. Zhong, M. Wang, H. Li, H. Zhang, Z. Lyu *et al.*, "You augment me: Exploring chatgpt-based data augmentation for semantic code search," in *2023 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 2023, pp. 14–25.

[12] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston, "Neural text generation with unlikelihood training," *arXiv preprint arXiv:1908.04319*, 2019.

[13] C. Wilt, J. Thayer, and W. Ruml, "A comparison of greedy search algorithms," in *proceedings of the international symposium on combinatorial search*, vol. 1, no. 1, 2010, pp. 129–136.

[14] V. Fernandez-Viagas, J. M. Valente, and J. M. Framinan, "Iterated-greedy-based algorithms with beam search initialization for the permutation flowshop to minimise total tardiness," *Expert Systems with Applications*, vol. 94, pp. 58–69, 2018.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need.(nips), 2017," *arXiv preprint arXiv:1706.03762*, vol. 10, p. S0140525X16001837, 2017.

[16] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[17] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[18] M. Thelwall, "Evaluating research quality with large language models: An analysis of chatgpt's effectiveness with different settings and inputs," *arXiv preprint arXiv:2408.06752*, 2024.