UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Can Large Language Models Replace Human in Speech Analysis?

Author:
Pol GARETA

Supervisor:
Santi SEGUÍ
Carolina MARTINEZ

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

January 17, 2024

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Can Large Language Models Replace Human in Speech Analysis?

by Pol GARETA

This thesis delves into the rapidly growing domain of Large Language Models (LLMs) and examines their relevance in the insurance sector, specifically focusing on their use in speech analysis to evaluate service quality. With the rapid escalation in the popularity of LLMs, we have the opportunity to analyze their practical use, focusing on Generali Seguros' customer service operations. This research is based on a partnership with Generali Seguros, which provided valuable access to audio recordings of their customer service calls and the associated evaluation templates used for assessing their teleoperators.

The core objective is to investigate the potential and real-world applications of LLMs in analyzing and evaluating the quality of service provided by Generali's teleoperators. To facilitate this, the study utilizes a secure and confidential environment provided by AWS, selecting commercially available models for analysis. The approach begins with converting the audio calls into Spanish text through an audio-to-text model, followed by improvements to this transcription method. Next, the study evaluates a baseline LLM that supports multiple languages and allows for fine tuning.

A significant aspect of this research includes addressing the challenges inherent in LLMs, such as their tendency towards 'inventing' responses and providing vague answers. Efforts to mitigate these issues involve both employing the baseline model in English —anticipating better performance due to its primarily English training—and implements strategies to enhance its effectiveness. Additionally, fine-tuning of the model is conducted, with the objective of specializing the model to required task.

Despite efforts to enhance the LLM, a notable finding of this study is the model's consistent failure to predict the minority group in the data, underscoring the limitations of current commercial models in fulfilling this specific evaluative function. The thesis concludes that, while LLMs show promise, they are yet to fully meet the demands of specialized tasks such as nuanced speech analysis in customer service settings. For transparency and further research, all codes used in this study are made available in a GitHub repository (Gareta, 2023).

Acknowledgements

First of all, I am extremely grateful to my supervisor at Generali Seguros, Carolina Martinez. Her leadership and mentorship have been unparalleled, making our collaboration an immensely rewarding and educational experience for me. I would also like to extend my thanks to Generali Seguros for providing the unique opportunity to work with real data.

My sincere gratitude goes to my academic supervisor at the University of Barcelona, Santi Seguí, for his invaluable guidance, innovative ideas, and the assurance that our efforts were progressing in the right direction.

Finally, I would like to thank my family, partner and friends for being there and for their kind support during this period.

Contents

Chapter 1

Introduction

1.1 Introduction to Large Language Models

In recent years, there has been a significant increase in interest surrounding Large Language Models (LLMs), especially after the introduction of new models like GPT-3 developed by OpenAI. This growing attention spans both academic and commercial sectors. Companies and researchers are increasingly attracted to the potential applications and innovations offered by these advanced models, indicating their growing importance in various domains. This trend is a critical milestone in the evolution of artificial intelligence and natural language processing and sets the stage for a deeper exploration of LLMs.

A LLM is an advanced type of machine learning model that specializes in processing and generating human language. LLMs are designed to understand, interpret, and produce text in a way that is coherent and contextually relevant. At the core of an LLM's functionality is its ability to learn from vast datasets of text, encompassing a wide range of human language nuances, including grammar, syntax, context, and even style. Unlike traditional language models which are more limited in scope and ability, LLMs use deep learning techniques and neural network architectures to handle a diverse array of linguistic tasks with high proficiency. This capability stems from their extensive training on diverse language data, enabling them to mimic human-like understanding and expression.

The LLMs' design typically involves layers of interconnected nodes in a neural network that process and transmit information, which allows them to generate responses that are not only relevant but also sophisticated in their understanding of language subtleties. An important aspect of LLMs is their adaptability through fine-tuning, allowing them to be specialized for specific tasks or domains, enhancing their utility and accuracy for a range of applications. This adaptability and sophistication make LLMs powerful tools in natural language processing, revolutionizing how we interact with and leverage machine-generated language.

1.1.1 Evolution and Technological Progress of LLMs

The journey LLMs is a fascinating tale of rapid advancement and innovation in the field of artificial intelligence and natural language processing. This evolution begins with the early, rudimentary models of the 1960s, such as ELIZA and PARRY (ELIZA, n.d.), which could mimic human conversation to a limited extent. These initial models establish the groundwork for future developments, despite their simplistic nature and reliance on pre-defined scripts.

As computational power increased and research in AI and machine learning deepened, more sophisticated models emerged. The introduction of statistical models in the late 1980s and 1990s marked a significant shift from rule-based systems to those capable of learning from large datasets. This shift was further accelerated with the advent of neural network-based models in the early 2000s, which brought a deeper understanding of language nuances.

The real game-changer came with the development of models like Google's BERT and OpenAI's GPT series. BERT (Bidirectional Encoder Representations from Transformers) revolutionized the way models understood the context within the text, allowing for more nuanced language processing. Meanwhile, the GPT series, especially GPT-3, pushed the boundaries in terms of scale and capability. GPT-3's ability to generate coherent and contextually relevant text is unparalleled, thanks to its 175 billion parameters and advanced training techniques.

These models owe their success to a combination of factors. Firstly, the advancement in neural network architectures, particularly the Transformer model, allows for more efficient processing of sequential data like language. Secondly, the vast amount of data available for training has enabled these models to learn a wide array of language patterns and styles. Finally, continuous improvements in hardware and cloud computing have made it feasible to train and run such large models.

1.1.2 Applications of LLMs

LLMs have been transformative in various fields, showcasing their versatility and impact. In conversational AI, they have enhanced the capabilities of chatbots and virtual assistants, providing more coherent and contextually relevant interactions, as seen in Google's Meena model. In content generation, LLMs like GPT-3 demonstrate their prowess by producing diverse textual content, including news articles and creative writing, revolutionizing the content creation process.

Another significant application is in sentiment analysis, where LLMs excel in extracting subjective information from texts, useful in customer feedback analysis and brand monitoring. This capability was notably demonstrated by BERT in analyzing sentiments in tweets related to the COVID-19 pandemic (Hawthorn, Weber, and Scholten, 2023). The field of machine translation has also seen remarkable advancements with LLMs, as exemplified by Google Translate's use of neural machine translation models for efficient and accurate translations across numerous languages.

LLMs have furthered developments in speech recognition, essential for virtual assistants and transcription services. They also play a crucial role in automated summarization, aiding in quickly understanding large volumes of text in domains like legal document analysis and academic research.

In education, LLMs contribute to personalized tutoring and generating educational content, enhancing learning experiences. Healthcare applications include processing medical documentation and assisting in preliminary diagnoses (Chang and Chang, 2023). In the financial sector, LLMs' ability to analyze market trends and interpret financial documents offers valuable insights for investment strategies. Finally, in the creative arts, LLMs are being used to generate music, poetry, and even assist in scriptwriting, showcasing their creative potential.

These diverse applications not only underscore the versatility of LLMs but also highlight their growing importance across various sectors, marking a significant shift in how technology and information are interacted with and processed in the digital age.

1.1.3 Challenges, Limitations, and Future Developments in LLMs

While LLMs represent a significant advancement in technology, they also confront substantial challenges and limitations. Ethical and privacy concerns are at the forefront, particularly regarding the use of sensitive data in training these models. This raises questions about intellectual property rights and consent. Additionally, LLMs often reflect biases present in their training data, leading to potential fairness issues in applications like hiring or lending decisions . The computational costs and environmental impact of training and deploying LLMs also present considerable challenges, especially for smaller organizations and academic institutions.

Looking to the future, LLMs are expected to evolve in ways that address these challenges. We might see advancements in autonomous models that can generate their own training data, leading to continuous self-improvement. Integration with other technologies such as virtual reality could open up new applications. A significant focus will likely be on developing methods for bias detection and mitigation, and on establishing ethical guidelines for LLM use. Improvements in computational efficiency are anticipated, which would reduce both the carbon footprint and the costs associated with LLMs. Enhanced customization options and broader accessibility are also expected, democratizing access to these powerful tools.

Despite the challenges, the future of LLMs appears promising, with potential advancements indicating a path forward that balances power with responsibility and ethical considerations.

1.2 Speech Analytics in Customer Service Sectors

Speech analytics is a technology that leverages advanced computational techniques to analyze spoken language. This powerful tool is used in various sectors, notably in customer service, to extract meaningful insights from voice interactions between customers and service representatives.

At its core, speech analytics involves the process of transcribing and analyzing spoken conversations to identify key phrases, sentiments, and patterns. It employs a combination of several technologies, including natural language processing (NLP), speech recognition, and machine learning. The primary objective is to understand better and interpret human speech in a way that is beneficial for businesses and organizations.

In a customer service context, speech analytics can review and analyze call recordings or real-time conversations with customers (Park and Gates, 2009). This analysis helps identify customer needs, preferences, and concerns. By capturing and analyzing the subtleties of language, tone, and speech patterns, companies can gain a deep understanding of customer sentiment and behavior.

One of the main applications of speech analytics is improving customer experience and service. By analyzing customer interactions, companies can identify areas of improvement in their service offerings, detect and address common customer issues, and train customer service representatives more effectively. Additionally, speech analytics can play a crucial role in compliance monitoring, ensuring that conversations adhere to legal and regulatory standards.

As technology advances, speech analytics is becoming increasingly sophisticated, offering more accurate and insightful analyses. This is largely due to improvements in AI and machine learning algorithms, which enable more nuanced understanding and interpretation of human speech.

Overall, speech analytics represents a significant advancement in how businesses can harness spoken language to enhance customer engagement, improve services, and make data-driven decisions.

1.3 Objectives and Scope

This thesis is centered around the application of commercial LLMs within a highly specialized context: the evaluation of customer service interactions in the insurance industry. The main objective of this research is to critically assess the capabilities of existing LLMs in handling the nuanced and complex nature of customer interactions specific to the insurance sector. This involves a detailed evaluation of these models' initial performance, understanding their baseline abilities in processing and interpreting the multifaceted layers of communication inherent in insurance-related dialogues.

A significant part of this study will be dedicated to investigating the potential modifications and fine-tuning techniques required to enhance the efficiency and accuracy of LLMs in this domain. This will include a thorough analysis of areas where current LLMs may lack in fully grasping and responding to the specific nuances of conversations related to insurance. The focus will be on developing strategies to bridge these gaps, thereby enhancing the models' comprehension and response capabilities. In line with this, the thesis will explore the feasibility of implementing these LLMs in a secure environment like Amazon Web Services (AWS), considering the stringent data privacy restrictions that govern the insurance industry. The choice of AWS as the platform for this study acknowledges the critical need for data security and privacy in handling sensitive customer information.

Furthermore, the research aims to provide insights into how these technological advancements can be leveraged to improve customer service quality in the insurance sector. By enhancing LLMs through targeted fine-tuning and modifications, the study seeks to push the boundaries of what these models can achieve, transforming them into more effective tools for analyzing and improving customer service interactions.

In summary, this thesis will not only offer an in-depth evaluation of the current state of LLM technology in the context of insurance customer service but also aims to advance the field by exploring and implementing improvements that could set new benchmarks for LLMs in customer interaction analysis.

1.4 Overview of the Collaboration with Generali Seguros

This thesis benefits significantly from the collaboration with Generali Seguros, offering a practical and real-world context for the application of LLMs. The focus of this partnership lies in analyzing voice call data collected between July and September 2023. The primary objective is to assess and enhance the efficiency of LLMs in evaluating customer service interactions, specifically comparing them against Generali Seguros' current manual evaluation methods. These manual methods, while traditional, have limitations in terms of time consumption and scalability, making the potential improvements through LLM application particularly valuable.

A crucial element of this collaboration is the use of a detailed evaluation template provided by Generali Seguros. This template comprises 24 items, each representing different criteria used in assessing the quality of customer service calls. The results generated by the LLMs will be compared against evaluations done using this template, offering a concrete measure for evaluating the effectiveness and accuracy of the LLMs in mimicking human evaluation processes.

Additionally, the nature of the data involved in this study - voice calls from insurance customers - requires stringent adherence to data protection and privacy standards. As a result, the thesis project will utilize AWS for data processing and analysis. AWS provides a secure and robust platform, ensuring that all data handling complies with the necessary privacy and security regulations. This approach not only aligns with legal and ethical standards but also offers a reliable and efficient means of managing the data of this study.

Through this collaboration, the thesis aims to provide valuable insights into the practical applications of LLMs in the insurance industry, demonstrating the potential benefits and improvements over traditional manual evaluation methods. The partnership with Generali Seguros is instrumental in grounding the research in a real-world scenario, thereby enhancing the relevance and applicability of the findings.

Chapter 2

Methodology

2.1 Data Collection

The dataset used in this Master's thesis is comprised of voice call recordings from Generali Seguros. A total of 88 calls were recorded between July and September 2023, which have been utilized to conduct this study. These calls, collected from the Generali Seguros call center, encompass various customer service scenarios, offering a valuable dataset for in-depth analysis in this study.

Regarding the technical specifications, the audio files are characterized by a sampling rate of 8000 Hz, suitable for effectively capturing human speech. The cumulative duration of these calls is approximately 511 minutes and 50 seconds. The lengths of the individual calls vary significantly, with the longest being 9 minutes and 54 seconds and the shortest at 2 minutes and 3 seconds. The average call duration is approximately 5 minutes and 48 seconds.

The overall audio quality in these recordings is generally clear, particularly from the side of Generali Seguros' representatives. However, there are notable variations in clarity in certain sections of the calls, especially those from the customers. These variations present unique challenges in speech analysis, which are crucial to address in the study.

2.2 Data Preprocessing

The data preprocessing stage was pivotal in transforming raw audio files from Generali Seguros into a format that could be analyzed by text-based LLMs. This process began with the secure download of audio files from the Amazon S3 server, which was carefully managed within the SageMaker Studio environment to ensure the confidentiality of the data.

In order to utilize text-to-text LLMs effectively, it was crucial to transcribe all voice calls into text. For this task, the `whisper-large-v2` model within AWS SageMaker was employed. This model is part of the Whisper series, renowned for its efficient speech recognition capabilities. Whisper is an open-source, multi-lingual speech recognition system that is available on Hugging Face, a popular platform for machine learning models. The `whisper-large-v2` model requires specific audio input conditions: each audio segment should not exceed 30 seconds, must be sampled at 16kHz, and should be in `.wav` format. To comply with these requirements, each audio file was split into 30-second segments and processed individually. The transcribed text from these segments was then compiled into a single `.txt` file, with each section labeled with the corresponding audio ID.

The `whisper-large-v2` was selected over other variants such as `whisper` and `whisper-large` for its enhanced accuracy and efficiency, which aligned with the objectives of our project. Interestingly, although 65% of the training data for this model is in English, it exhibited strong performance in transcribing Spanish, our primary language data. The model's ability to automatically detect and transcribe the language of the audio significantly streamlined our workflow.

However, we faced a challenge in that the output text did not distinguish between different speakers in the conversations. Ideally, an extension like pyannote-audio which can identify individual speakers, would be advantageous. However, due to the constraints of operating in a commercial AWS environment, we could not use this extension as it requires external licensing not available in our AWS service framework. This limitation hindered our ability to analyze each speaker's contributions in the calls separately.

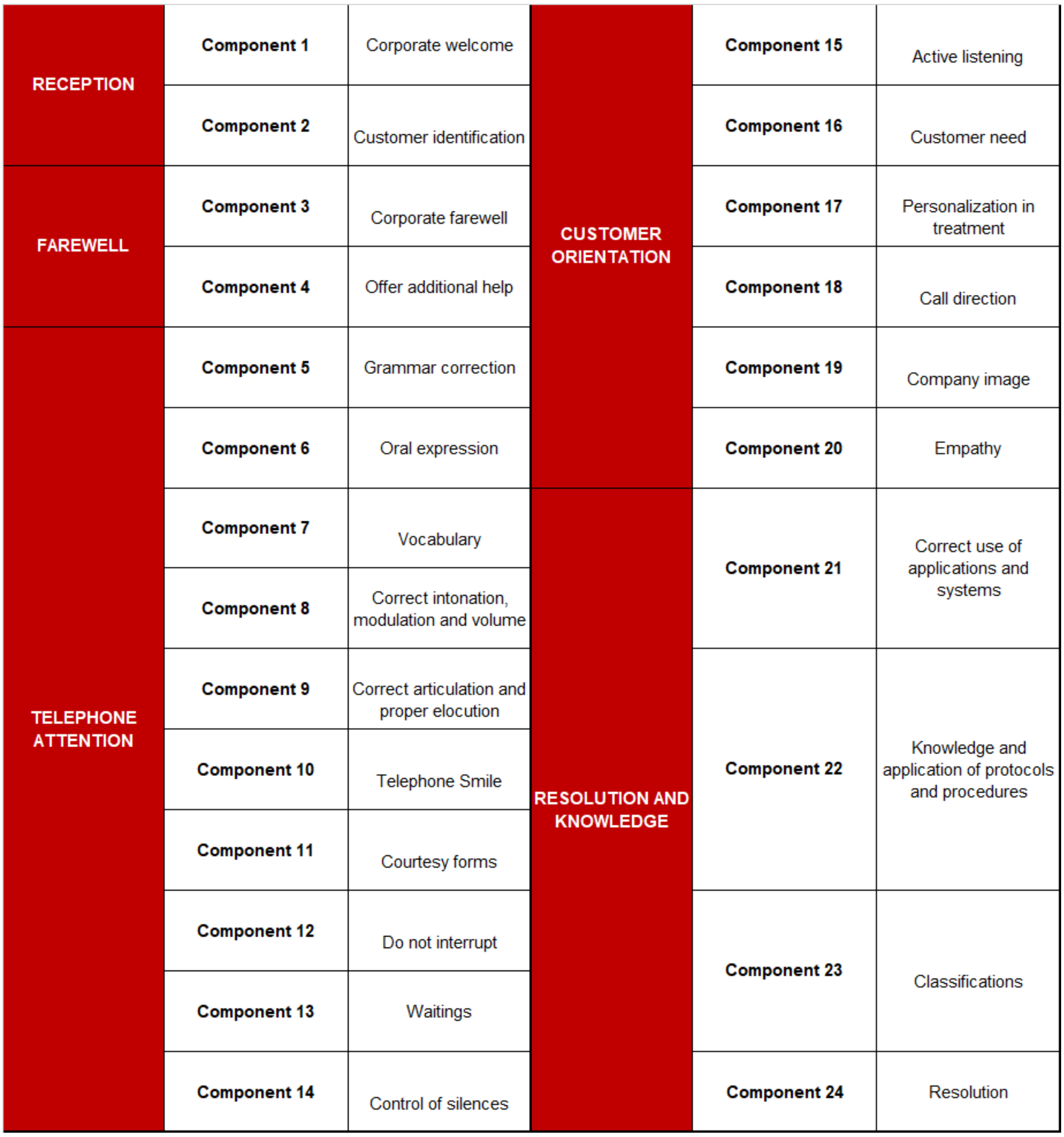
In addition, some errors appeared in the transcriptions, with certain words being repeated multiple times. To address this, we implemented a filter to remove any word that appeared repeated within a span of five words. This step was crucial for maintaining the accuracy and clarity of the transcribed data.

2.3 Evaluation Template

The evaluation of LLMs have been made using a carefully designed template to manually review calls from Generali Seguros. This template is a critical tool, divided into five main groups reflecting distinct phases of a customer service call: reception, farewell, telephone attention, customer orientation, and resolution and knowledge. These groups encompass a total of 24 subsections (Figure 2.1), each representing a feature for classification in the study. The specific criteria for assessing each feature are exemplified in an annex, which includes a list of actions or indicators that define the fulfillment of each feature. The template's importance is underscored by its direct influence on the remuneration of teleoperators at Generali Seguros, with their scores impacting their compensation.

An in-depth analysis of 88 corrected calls from Generali Seguros revealed an unbalanced dataset, with a predominance of positive assessments (1s) over negative ones (0s), marked by approximately a 79% ratio. This imbalance presents a unique challenge for the study. Additionally, a thorough component analysis of the calls was undertaken to gain a deeper understanding of the assessment criteria and their practical application. These detailed insights, including specifics from the component analysis, are illustrated in Figure 2.2.

The analysis represented in Figure 2.2 highlights a notable pattern, particularly in component 5, which consistently registers 100% in positive assessments. This trend suggests a potential bias in the data, where a model trained on this dataset might default to assigning positive scores for this feature, rather than learning nuanced distinctions. This observation is crucial, as component 5 pertains to grammatical correctness, a standard expectation in customer service calls. However, the study also identifies challenges in predicting certain features like `telephone smile` and `active listening` based solely on textual analysis. These features inherently require an understanding of tone and spoken nuances, which may not be fully captured in a text-based LLM analysis.



RECEPTION	Component 1	Corporate welcome	CUSTOMER ORIENTATION	Component 15	Active listening
	Component 2	Customer identification		Component 16	Customer need
FAREWELL	Component 3	Corporate farewell		Component 17	Personalization in treatment
	Component 4	Offer additional help		Component 18	Call direction
TELEPHONE ATTENTION	Component 5	Grammar correction		Component 19	Company image
	Component 6	Oral expression		Component 20	Empathy
	Component 7	Vocabulary	RESOLUTION AND KNOWLEDGE	Component 21	Correct use of applications and systems
	Component 8	Correct intonation, modulation and volume			
	Component 9	Correct articulation and proper elocution		Component 22	Knowledge and application of protocols and procedures
	Component 10	Telephone Smile			
	Component 11	Courtesy forms			
	Component 12	Do not interrupt		Component 23	Classifications
	Component 13	Waitings			
	Component 14	Control of silences		Component 24	Resolution

FIGURE 2.1: Equivalence Component-Item

2.4 Evaluation Metrics for LLM Performance

In this thesis, a comprehensive evaluation of LLMs will be conducted using a range of metrics, each providing insights into different facets of model performance within the context of an unbalanced dataset:

- **Precision:** Precision is a key metric that measures the accuracy of the model in identifying true positive outcomes. It calculates the proportion of positive identifications that are actually correct. High precision ensures that the model accurately identifies relevant cases without misclassifying them. This accuracy is essential to avoid misunderstandings or the dissemination of incorrect information. The formula for calculating precision is:

$$Precision = \frac{TP}{(TP + FP)}$$

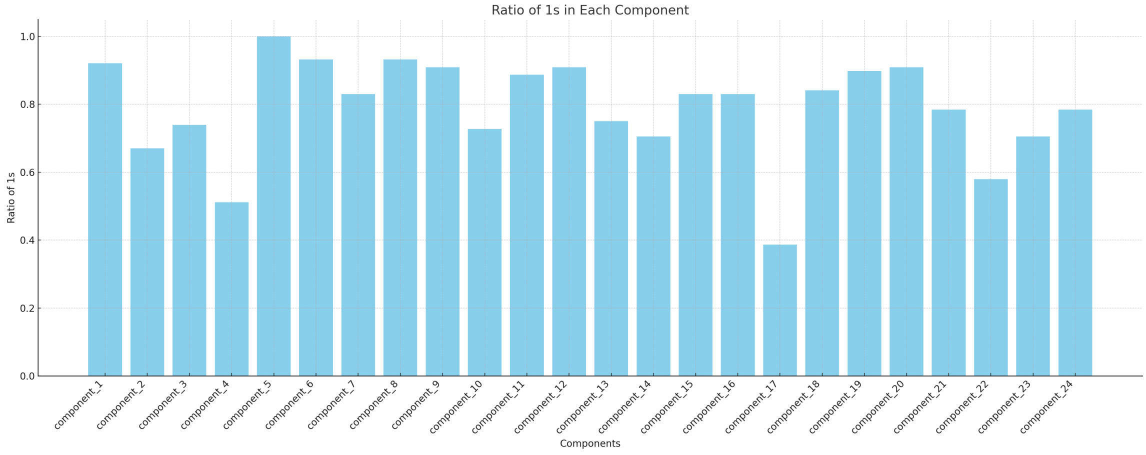


FIGURE 2.2: Ratios of 1s for each component

In this formula, TP represents the number of true positives, and FP stands for the number of false positives. A high precision score indicates that the model is effective in correctly identifying positive cases, making it a reliable tool for decision-making processes.

- **Recall:** Recall is an essential metric that measures the model's ability to capture all actual positive cases in the dataset. It determines the proportion of actual positives that the model correctly identifies. The formula for calculating recall is:

$$Recall = \frac{TP}{(TP + FN)}$$

In this formula, TP represents the number of true positives, and FN stands for the number of false negatives. A high recall score indicates that the model is effectively capturing all the relevant cases.

- **F1 Score:** The F1 Score is a crucial metric that combines precision and recall into one single measure, providing a balanced view of a classifier's performance. It is especially useful when the importance of precision and recall is roughly equivalent. A high F1 Score indicates that the classifier is both accurate (precision) and comprehensive (recall) in its identification of true positives. On the other hand, a low F1 Score suggests that the classifier is either not precise, not comprehensive, or both. The formula for calculating the F1 Score is:

$$F1\ Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

- **Matthews Correlation Coefficient (MCC):** The MCC offers a comprehensive evaluation of a classifier's performance, taking into account all quadrants of the confusion matrix (Chicco and Jurman, 2020). It is calculated using the formula:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

In this formula, TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively. Given the unbalanced nature of our dataset, with a predominance of positive assessments, an MCC value significantly higher than zero would indicate the model's robustness and its suitability for analyzing complex, real-world customer service conversations. The MCC is particularly informative in scenarios with unbalanced data, providing a more reliable measure of performance than accuracy alone.

Each of these metrics will be meticulously calculated for the 24 features delineated in the evaluation template. The individual feature scores will then be aggregated to yield an overarching measure of the model's effectiveness. By employing this suite of metrics, the thesis aims to present a nuanced and detailed assessment of LLM performance, taking into account the unique challenges posed by the dataset's imbalance and the critical nature of accurate communication in the insurance sector. This multifaceted approach ensures a holistic understanding of the model's capabilities and guides the identification of targeted improvements for enhancing LLM application in customer service evaluation.

Chapter 3

Implementation and Results

3.1 Model Selection

In the pursuit of this thesis, after the dataset was established and all conversations were meticulously transcribed into Spanish, the imperative next step was the selection of the most suitable LLM for performance evaluation within the operational confines of AWS's services. This decision was heavily influenced by the limitations highlighted in Section 1.1, which noted the general lack of specific functionality among commercial LLMs. Despite these challenges, the possibility of fine-tuning models to specific tasks presented a way to overcome these issues. Therefore, we focused on choosing a model that allowed fine tuning and is able to process queries in Spanish, not only in English, as the most of the model we found out.

The model that emerged as the most fitting candidate was Llama-2-Chat 7B, a derivative fine-tuned for dialogue-driven scenarios from the Llama-2 base model. This selection was made over its larger counterparts, such as Llama-2 13B or 70B, which were not feasible options within the AWS service framework. Llama-2-Chat 7B strikes a balance, offering an optimal blend of size and functional capability, tailored to dialogue applications. It is part of an emerging breed of open-source LLMs, which have been explicitly engineered to exhibit advanced conversational abilities.

The foundational training of the base model, Llama-2, was conducted unsupervised on a vast corpus of publicly accessible text sources, encompassing 2 trillion tokens. This expansive training set assures a broad linguistic and knowledge base, although it is important to note that it does not contain Meta user data and is primarily tailored for the English language. The fine-tuning of Llama-2-Chat 7B was executed through Reinforcement Learning from Human Feedback (RLHF), employing advanced methodologies such as Importance Sampling and Proximal Policy Optimization. This process was meticulously calibrated to refine the model's response quality, enhancing its utility and safety in mirroring human conversational patterns.

Llama-2-Chat models are specifically designed for dialogue and interaction-focused tasks, and their fine-tuning process enhances their ability to handle conversational contexts. This specialization makes Llama-2-Chat an optimal choice for our project, which demands natural language generation in dialogue systems. Its adaptability and targeted training for conversational scenarios make it well-suited for this thesis, addressing the intricate details of customer service interactions in the insurance industry.

3.1.1 Baseline Model Implementation

To initiate the practical application of a LLM within AWS, the initial step involved deploying an unmodified model to serve as a baseline for subsequent analyses. Interaction with this baseline model involved using well-designed queries, set with specific parameters to draw out the most informative and relevant responses. These parameters included the max_length_token, dictating the maximum token count for each model-generated response, and two pivotal parameters, temperature and top_p, which play a significant role in the variability and precision of the output. The temperature parameter, by controlling the randomness of the prediction, can lead to either more creative or more conservative outputs. To strike a balance between inventiveness and accuracy, the temperature was tuned to 0.6, encouraging a moderate level of unpredictability in the responses. The top_p parameter, on the other hand, helps in narrowing down the choice of words to a defined probability set, and was set to 0.9 to ensure that the model's responses were drawn from a broad yet probable range of vocabulary.

For the purpose of this study, a structured query skeleton was formulated for each assessable feature within the evaluation template, guiding the LLM to assign a binary '1' or '0', indicative of the fulfillment or lack of the specified criteria.

During the output extraction process, it was noted that the LLM often prefaced its binary responses with brief justifications. To capture the essence of these explanations without being overwhelmed by verbosity, the maximum token count was limited to 500, with a focused attention on the initial 300 tokens where the binary decision typically surfaced. In instances where neither a '1' nor '0' could be discerned, the query was repeated until we obtain a answer with '1' or '0'. To enhance the reliability of the results, each query was evaluated three times. The outcome that appeared most frequently in these trials was considered the final decision for that specific feature. This method of repeating the query three times was designed to mitigate the tendency of LLMs to sometimes generate fabricated responses—a challenge well-documented in LLM research (Azaria and Mitchell, 2023)—ensuring a more robust and dependable evaluation process.

As a result of this systematic querying and evaluation approach, we created a binary vector for each audio ID, reflecting the model's assessment based on the evaluation criteria. Importantly, the LLM processed each query independently, without considering any context from previous queries. This necessitated that each query be complete in itself, containing the full transcription of the respective call. This isolated treatment of queries is due to the LLM's design, which is not configured to process sequential or related data across different prompts.

3.1.2 Results

After the model completed its predictions for the entire set of audio files, the outcomes were compiled into a CSV file to facilitate a comprehensive analysis. An initial examination of the data revealed that a majority of the model's predictions —approximately 95%—were classified as '1s' Table 3.1, demonstrating a significant inclination towards positive evaluations across most of the components. This trend was consistent, with notable exceptions in components 7 and 17, which pertain to vocabulary usage and personalized treatment, respectively. These components, as depicted in Figure 3.1, displayed predictions with less than 80% of '1s', indicating a more balanced distribution of the model's assessments.

"1"	**"0"**
95.03%	4.97%
2007	105

TABLE 3.1: Output distribution

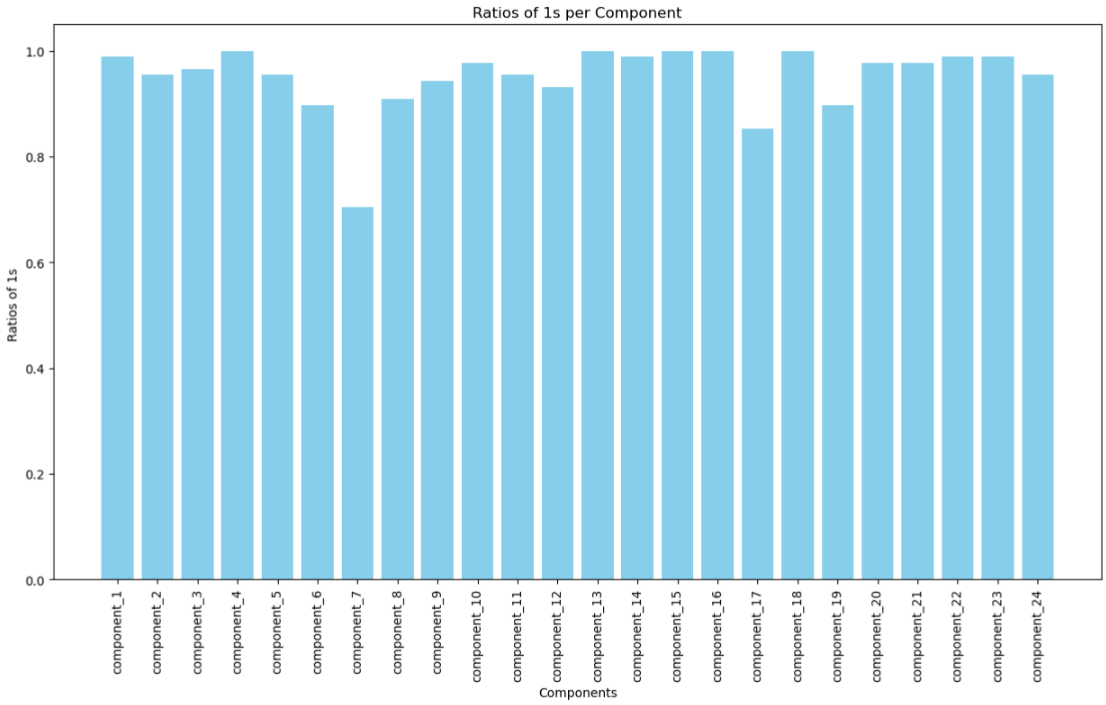


FIGURE 3.1: Ratios of 1s for each component

To critically assess the model's performance, the four metrics outlined in the section 2.4 —precision, recall, F1 score, and the MCC— were employed. The aggregate performance captured in Table 3.2 suggested a high level of precision and recall, as well as F1 scores, hinting at the model's proficiency in identifying true positive outcomes. The MCC, however, presented a contrasting perspective. With its values close to zero, this metric highlighted the model's struggle in accurately classifying negative instances. This indicates a significant gap between the model's ability to predict positive outcomes and its effectiveness in identifying true negatives.

Precision	Recall	F1-Score	MCC
0.7917	0.9544	0.856	0.0045

TABLE 3.2: Global performance metrics

A more detailed evaluation was carried out for each component, with the results detailed in Table 3.3. This analysis showed that components with a higher number of '1s', like components 5 and 6, tended to have better precision. However, despite these high precision scores, the overall effectiveness of the classifier was akin to random guessing. This is indicated by the MCC values, which were mostly near zero. This result confirms the limitations of the classifier, highlighting its inability to effectively differentiate between classes.

This analysis not only highlight the complexity of applying LLMs to specific tasks such as evaluating insurance customer service interactions but also highlights the importance of utilizing a range of metrics to capture a complete picture of model performance.

Component	Precision	Recall	F1-Score	MCC
component_1	0.9195	0.9877	0.9524	-0.0315
component_2	0.6786	0.9661	0.7972	0.0791
component_3	0.7294	0.9538	0.8267	-0.1118
component_4	0.5114	1.0000	0.6767	0.0000
component_5	1.0000	0.9545	0.9767	0.0000
component_6	0.9241	0.8902	0.9068	-0.0913
component_7	0.8065	0.6850	0.7407	-0.0948
component_8	0.9375	0.9146	0.9259	0.0713
component_9	0.9036	0.9375	0.9202	-0.0776
component_10	0.7209	0.9688	0.8267	-0.0934
component_11	0.8929	0.9615	0.9259	0.0938
component_12	0.9146	0.9375	0.9259	0.0713
component_13	0.7500	1.0000	0.8571	0.0000
component_14	0.7126	1.0000	0.8322	0.1656
component_15	0.8295	1.0000	0.9068	0.0000
component_16	0.8295	1.0000	0.9068	0.0000
component_17	0.4400	0.9706	0.6055	0.2646
component_18	0.8409	1.0000	0.9136	0.0000
component_19	0.9114	0.9114	0.9114	0.1336
component_20	0.9070	0.9750	0.9398	-0.0482
component_21	0.7791	0.9710	0.8645	-0.0800
component_22	0.5747	0.9804	0.7246	-0.0913
component_23	0.7011	0.9839	0.8188	-0.0694
component_24	0.7857	0.9565	0.8627	0.0181

TABLE 3.3: Updated Performance metrics for each component

3.2 Model Improvement

In this section, we aim to enhance the model's performance, building upon the baseline established with the raw, Spanish-transcribed model. The first improvement step involves testing the model with English translations of the original Spanish audio, based on that most of the Llama-2 training data was in English. This approach is expected to yield better results. The second step is model fine-tuning. This will include adjusting the model to align more closely with the specific characteristics of the dataset, with a focus on addressing the initially low MCC scores and the model's limitations in accurately classifying negative instances. The goal is to achieve a more balanced, accurate, and reliable performance from the model.

3.2.1 English Model

To assess our model's performance in English, it was essential to have all conversations transcribed into English. Consequently, we modified the 'whisper-large-v2' queries, originally designed to transcribe Spanish calls, to directly convert them into English transcriptions. A preliminary review of these transcriptions suggested that the model performed more effectively in English, avoiding the repetitive and nonsensical word patterns observed in the Spanish version.

Additionally, we translated all queries into English. This step seemed to set the stage for execution, but now entirely in English. Our initial hypothesis was that, given the predominance of English training data, the model would yield more accurate responses. Upon the first execution in this new setup, the code appeared to run faster. However, it became apparent that the model was simply mirroring our queries. Our queries often started with "The response should only be 1 or 0", leading the model to consistently include "1" in its responses, inaccurately suggesting successful results.

To address this, we revised our approach: an answer would only be considered valid if it contained exclusively "1" or "0". If both or neither were present, the query would be repeated. Despite these changes, the code entered an infinite loop. The model consistently failed to provide the correct single-digit response, never concluding its process. To counter this, we implemented a cap, limiting each query to a maximum of five attempts. If no unique response emerged after these attempts, we defaulted the output to "2", signifying the model's inability to respond.

Our initial method to enhance output robustness — repeating each query thrice and selecting the most common response — proved ineffective. We altered this strategy, accepting an answer only if it was consistent across all three output vectors. Otherwise, the final output for that item would be marked as "3". It is important to note that while the fundamental issue remains a binary "1" or "0" problem, outputs labeled "2" or "3" are considered invalid and discarded. This dual-error output system was introduced to provide more detailed insights into the model's performance.

To be able to compare the performance between the model tested in English and Spanish, we repeated both experiments with this new conditions.

Results

After completing all interactions with the model, we will now focus on examining the aspect of proportionality. A detailed analysis of the data distribution is provided in Table 3.4, which presents the proportions of '1', '0', and non-valid outputs. It is important to note that the execution in English displayed a higher speed compared to that in Spanish.

Language	"1"	"0"	No Valid Outputs	
			"2"	"3"
English	0.64	0.002	0.060	0.298
Spanish	0.80	0.025	0.015	0.16

TABLE 3.4: Proportionality of Data

Based on the comparison, it's clear that about 30% of the model's outputs in English queries were discarded due to inconsistencies in at least one of the three outputs. In contrast, the Spanish version of the model saw around 16% of non-valid outputs. Notably, in both cases, the proportion of '0' outputs is quite low. Next, we will delve into the analysis of the four key metrics, Table 3.7.

Language	Precision	Recall	F1-Score	MCC
English	0.78	0.99	0.87	-0.002
Spanish	0.80	0.97	0.83	0.014

TABLE 3.5: Global performance metrics

In the next section, we will conduct a comprehensive analysis of the results. However, it is evident that we observe similar behavior and insignificant improvements compared to the baseline model.

3.2.2 Model Fine Tuning

In the previous section, we observed that even when using the most popular language in the trained data, the model's performance remained low, with the MCC approaching 0, which is as having a random classifier. Our next step, as previously discussed, is to train a model with specific data to create a task-specific model. In our case, this model aims to answer the evaluation template based on call conversations.

For our model, LLama-2-7b Chat, the fine-tuning process requires a JSON file with different items, each following a specific format. Each item consists of "### Instruction," followed by the query and the transcription of the call, and then "### Response," where we provide the response, being so concise, just adding "The evaluation for this item is X".

The initial step involves constructing our training data. Given that the model tends to produce fewer invalid outputs in Spanish, we decided to train it using Spanish data, as there was no significant difference in the four main metrics between languages. Therefore, we added the same instructions to the queries as in the previous experiments. First, we specified whether we wanted to answer '1' or '0', based on the subsequent items. Finally, we included the conversation transcription. In the response field, we obtained the answer from the template. We repeated this process for all items across all files, resulting in a total of 2112 items for training.

Once the model is deployed, we followed the same evaluation steps as in Section 3.2. We anticipated better results since we were evaluating the model with the same data it was trained on.

Upon analyzing the answers, we observed improvements compared to the previous vectors. With this trained model, the answers consistently followed the structure of the training data, beginning with "The evaluation for this item is X," followed by additional text. The extra text sometimes included justifications but also extracted information from the conversation. Unlike the previous model, where we obtained the entire 500 characters of the answer, we now extracted only the first 50 characters, which typically contained the essential information. This adjustment was possible because all responses followed the same structured format.

However, it became apparent that the model always generated '1' as its output for every query. This behavior seemed to be rooted in the training data, which had an 80% ratio of '1s'. To address this, we applied a preliminary filter to the training data, removing samples with more than 19 '1s' out of 24 items. While this step resulted in the removal of 42 items from the dataset, it effectively reduced the '1s' ratio to 65%. Subsequently, we fine-tuned the model using this modified dataset.

Results

The results for the fine-tuned model offer a detailed evaluation of the LLM's effectiveness, with a notable slowdown in processing speed as a key factor in its performance. This slowdown is important to consider, as it may affect the model's usability in real-time scenarios. Like in previous phases, the first step is to look at the data's proportionality, and this information is provided in Table 3.6.

"1"	"0"	No Valid Outputs	
		"2"	"3"
0.53	0.02	0.00	0.45

TABLE 3.6: Proportionality of Data

From the proportionality table, a noticeable increase in the number of invalid responses becomes apparent. The dataset exclusively contains ’3s’ for invalid parameters, suggesting that within a maximum of five attempts, the model was always able to produce an answer. This absence of ’2s’ indicates that while the model does not fail to provide a response, there exists a heightened level of unpredictability in its outputs, with approximately 40% of the items displaying variability. Furthermore, while still low, there is a slight upgrade in ratio of ’0s’ when compared to previous models, suggesting a modest improvement in the diversity of the model’s responses.

Moving to the assessment of the evaluation metrics, it is observed in Table 3.7, the precision, recall, and F1 Score continue to register high values. Nonetheless, these metrics alone do not offer a conclusive representation of the model’s ability to classify effectively, as evidenced by the consistently low MCC. The MCC’s near-zero value suggests that the model’s ability to distinguish between positive and negative classes has not markedly improved. Despite the high values in the other metrics, the low MCC implies that the model’s predictive improvements are not substantial enough to classify it as a robust classifier.

Precision	**Recall**	**F1-Score**	**MCC**
0.80	0.97	0.88	0.031

TABLE 3.7: Global performance metrics

Chapter 4

Results Analysis and Discussion

Following the implementation of our planned improvements steps, the analysis and discussion of the results offer a chance for an in-depth comparative evaluation. Our initial evaluation with the first model, specifically tailored for Spanish, displayed minimal differences when contrasted with the latest, theoretically more advanced model as is shown in Table 4.1. However, this negligible variance, noted across a dataset comprising 2112 items, lacks statistical significance. This indicates a persistent challenge faced by all models in correctly identifying negative values, a vital aspect in our research context.

Model	Precision	Recall	F1-Score	MCC
Baseline	0.79	0.95	0.86	0.0045
English	0.78	0.99	0.87	-0.002
Spanish	0.80	0.97	0.83	0.014
Fine Tuned	0.80	0.97	0.88	0.031

TABLE 4.1: Global performance metrics for all the models

To provide a deeper understanding of these values, we have included in Table 4.2 the confusion matrices for all the methods. These matrices clearly illustrate the significant disparity between True Positives (TP) and True Negatives (TN), highlighting the unbalanced nature of the dataset. Of particular note is the performance of the model tailored for English. This model shows the poorest performance in our evaluations, with it correctly labeling only one item as "0".

Model	TP	FP	TN	FN
Baseline	1588	419	24	81
English	1060	290	1	4
Spanish	1338	344	13	42
Fine Tuned	904	233	10	26

TABLE 4.2: Confusion Matrix for all the models

Significant progress was made in terms of response rate optimization. The initial model exhibited prolonged processing times for all samples, largely due to the absence of a limit on query attempts before obtaining a response. Initially, we accepted only those responses with "1" or "0" that appeared within the first 300 characters. Implementing a filter to exclude samples displaying both "1" and "0" simultaneously resulted in improved processing speed and efficiency, with minimal compromise on response quality. This strategy of selecting the most frequent value from three attempts inadvertently masked an underlying issue with Large Language Models:

their tendency to generate responses aimed at minimizing predicted errors, even in the absence of definitive answers (Table 4.3).

Despite the model's primary training in English, no significant differences in performance were observed, with the primary changes being in response speed and an increase in discrepancies in answers.

Model	**"1"**	**"0"**	**No Valid Outputs**	
			"2"	**"3"**
Baseline	0.95	0.50	-	-
English	0.64	0.002	0.060	0.298
Spanish	0.80	0.025	0.015	0.16
Fine Tuned	0.52	0.03	-	0.45

TABLE 4.3: Proportionality of Data of all models

Our ultimate refinement effort involved training the model with specific data. We encountered limitations due to the restricted availability of data and the inability to utilize augmentation techniques, leaving us with origial 88 audio samples for analysis. These samples exhibited inherent biases, leading to a classifier skewed towards consistently predicting "1," thus diminishing its effectiveness.

When we attempted to reduce biases by filtering the data, we lost half of the dataset, making it insufficient for effective training. Nevertheless, training with this refined dataset did yield some improvements, not in the classification problem, but improve manner in which the models responded to queries, it limits the output to the one requested. Something that since the beginning, we have been asking the model to be concise with the output, just to answer "1" and "0." However, as it is a LLM, its behavior is limited.

From the metrics and results analysis, it's crucial to recognize that certain evaluation aspects, especially those related to telephone attention like component 10 ('Telephone Smile') and component 12 ('Control of Silences'), inherently carry complexities when assessed through transcriptions alone. Initially, the decision was to keep these items active until we observed any significant variance in the model's predictions. However, the absence of such variances prompted us to retain all values, as there was insufficient evidence to conclude that the model could accurately and consistently predict these specific metrics. This approach underscores the model's unpredictable and inconsistent performance and underlines the inherent difficulties in deploying LLMs for tasks demanding a certain level of interpretation and understanding.

Chapter 5

Conclusion

In this Master's Thesis, we conducted a comprehensive evaluation of various approaches to enhance the performance of a commercial LLM. We began with the fundamental implementation of querying the model and gradually progressed to more sophisticated approaches involving fine-tuning with specialized data. A central objective of this research was to explore the feasibility of employing current commercial LLMs as replacements for human analysis in speech evaluations, particularly focusing on customer service interactions within an insurance company setting, utilizing a meticulously structured template. Despite employing a variety of techniques and improvements, our findings indicated that, given the limitations of current resources and technological capabilities, LLMs are not yet sufficiently equipped to act as viable substitutes. While LLMs have experienced rapid advancements recently, they still lack certain critical capabilities required for specific tasks.

Throughout this investigative process, significant improvements were anticipated at each stage of model modification. However, each step unveiled new and unexpected challenges. One of the initial obstacles was applying LLMs effectively in real-world, commercial settings such as AWS. This environment presented specific constraints, including licensing limitations and a limited selection of models capable of handling multilingual tasks and intricate fine-tuning processes. An additional and significant insight gained from our research was the inherent complexity and substantial expense associated with the acquisition of new, high-quality data. We found that even when data was accessible, its biased nature frequently posed considerable challenges, significantly impeding the generation of accurate predictions and evaluations.

Our experimental efforts with the model in English resulted in only marginal improvements. The audio-to-text model showed more promise, yet the anticipated breakthroughs in the fine-tuning process did not materialize as expected, leading to notable shortcomings in the evaluation process of the commercial LLM.

A major limitation encountered in our study was the choice of transcription and LLM models. The inability of transcription models to identify individual speakers or to ascertain the precise timings of their speech segments adversely impacted the overall performance. Additionally, the inherent design of LLMs, which processes each query independently, presented a significant challenge, particularly with lengthier queries that tended to result in less precise outputs. Advanced models like GPT-4, capable of processing entire conversations before evaluating against specific criteria, could potentially offer a more accurate and comprehensive analysis.

A critical problem we encountered in our research was the 'invention' tendency of LLMs (Azaria and Mitchell, 2023), wherein the outputs occasionally lacked veracity. This issue had a significant impact on the reliability of the model's outputs and diminished the explainability of its evaluations for each criterion. This shortcoming becomes particularly worrisome in scenarios where model outputs have direct implications on economic decisions.

To mitigate this issue, we implemented a condition that only accepted samples which showed consistent output across three separate attempts. This approach aimed to enhance the reliability of the results by favoring consistency, under the assumption that repeated agreement on an evaluation would indicate a higher likelihood of accuracy. However, this strategy inadvertently led to a considerable reduction in the number of usable samples. Many responses did not meet the criterion of consistency, resulting in their exclusion from the final analysis. This loss of data samples presented a new challenge: it narrowed the scope of our analysis and potentially impacted the robustness of our findings. By prioritizing consistency to counteract the invention tendency of LLMs, we faced a trade-off between the quantity of data and the perceived quality or reliability of the model's outputs, highlighting another layer of complexity in the application of LLMs for specific tasks such as speech evaluation in customer service scenarios.

Furthermore, LLMs' tendency to generate extensive text often contradicted our need for concise answers. This aspect, while beneficial in some contexts, was a impediment in our study where brevity and accuracy were crucial.

Additionally, we encountered difficulties in measuring LLM performance. The extraction of precise answers was challenging, and traditional metrics like precision, recall, or F1-score did not fully explain the model's behavior. We had to incorporate an additional metric, the MCC, highlighting the importance of focusing on the significance of outputs. In our case, we needed to consider both positive and negative predictions, emphasizing the need for a balanced approach in evaluating the performance of LLMs. This revelation about the necessity of a multifaceted metric system, like including MCC alongside traditional metrics, underscores the complexity of accurately assessing LLM capabilities. It's crucial to discern not just the overall accuracy but also how well the model distinguishes between different types of responses, especially in scenarios where both positive and negative predictions carry significant weight (Chicco and Jurman, 2020).

In conclusion, this thesis has illuminated the challenges and high costs of acquiring quality data for LLMs, particularly unsupervised models, while also highlighting the differences between academic research and real-world applications. Despite some setbacks, our findings emphasize the potential of LLMs, provided they are harnessed correctly with targeted fine-tuning and comprehensive performance evaluation. Our project, though not fully achieving its goals, has offered valuable insights into the capabilities and limitations of current LLM technology, contributing to future advancements in natural language processing and its practical applications.

Chapter 6

Future directions

The field of speech analytics, especially when enriched by advancements in deep learning and machine learning, offers lots of research opportunities that have the potential to transform various applications, including those in the insurance sector and call analysis.

A particularly promising area of future exploration is the analysis of silences within conversations. Gaining an understanding of the reasons and contexts for these pauses can yield valuable insights, which could be pivotal in refining communication strategies, potentially leading to substantial cost reductions for businesses. The integration of advanced technologies could facilitate the identification of patterns in the flow of conversation, providing a deeper understanding of communication effectiveness and efficiency.

Another promising direction for future research lies in enhancing the capability of technologies to generate detailed summaries of conversations. This feature would be especially beneficial in scenarios where rapid assimilation of information is critical, such as in customer service or during strategic decision-making processes.

Further developments in the performance of LLMs could also address some of the challenges identified in our research. More sophisticated and robust LLMs could potentially overcome limitations in specific task performance, thereby increasing their suitability for real-world applications.

Additionally, the use of these technologies in anonymizing data could offers a practical solution to the privacy and data sensitivity concerns highlighted during our study. As LLMs and related technologies continue to evolve, they are likely to provide increasingly advanced methods for data anonymization. This would balance the need for privacy with the ability to extract valuable insights from data.

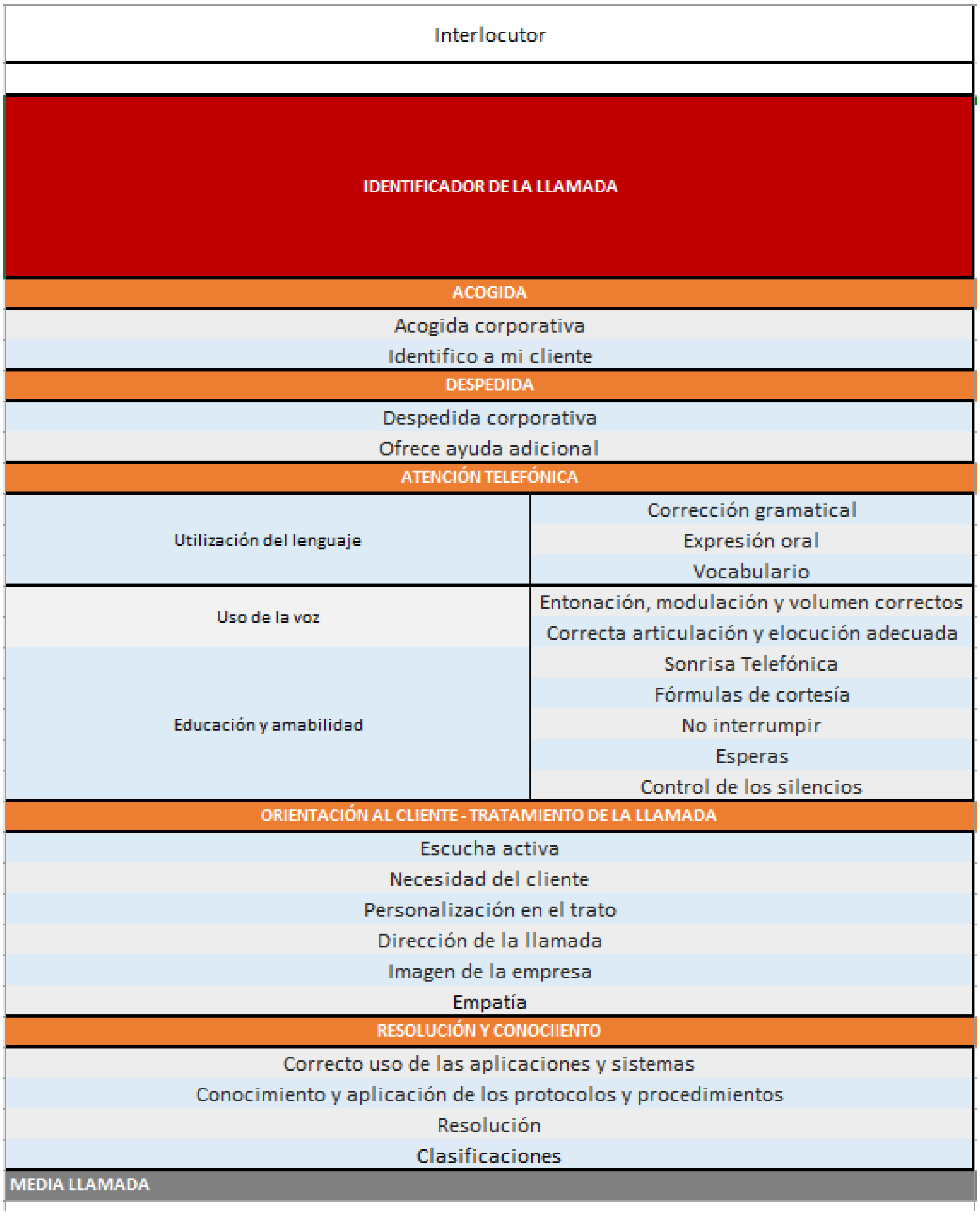
Moreover, there is potential for the application of speech analytics in enhancing customer experience and personalization. By analyzing speech patterns and content, businesses could tailor their services and interactions to better meet individual customer needs, leading to improved customer satisfaction and loyalty.

The application of speech analytics in automated and real-time feedback systems also presents exciting possibilities. Such systems could provide immediate insights and suggestions to customer service representatives during live calls, aiding in more effective communication and problem resolution.

In summary, the integration of advanced machine learning and deep learning technologies in the realm of speech analytics offers a bright future. It holds the promise of not only refining current processes but also opening new paths for research and practical applications, especially in sectors where effective communication and data analysis are crucial. This area of study stands at the forefront of technological innovation, with the potential to significantly impact various industries and improve numerous aspects of professional and everyday communication.

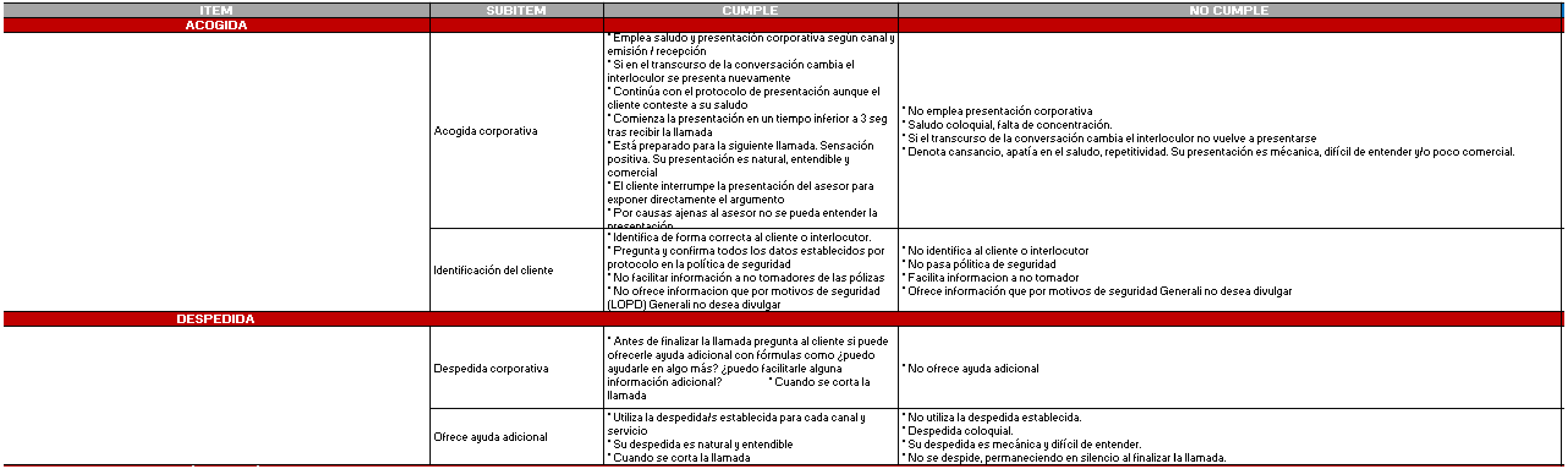
Appendix A

Evalutation Template



Interlocutor	
IDENTIFICADOR DE LA LLAMADA	
ACOGIDA	
Acogida corporativa	
Identifico a mi cliente	
DESPEDIDA	
Despedida corporativa	
Ofrece ayuda adicional	
ATENCIÓN TELEFÓNICA	
Utilización del lenguaje	Corrección gramatical
	Expresión oral
	Vocabulario
Uso de la voz	Entonación, modulación y volumen correctos
	Correcta articulación y elocución adecuada
Educación y amabilidad	Sonrisa Telefónica
	Fórmulas de cortesía
	No interrumpir
	Esperas
	Control de los silencios
ORIENTACIÓN AL CLIENTE - TRATAMIENTO DE LA LLAMADA	
Escucha activa	
Necesidad del cliente	
Personalización en el trato	
Dirección de la llamada	
Imagen de la empresa	
Empatía	
RESOLUCIÓN Y CONOCIIENTO	
Correcto uso de las aplicaciones y sistemas	
Conocimiento y aplicación de los protocolos y procedimientos	
Resolución	
Clasificaciones	
MEDIA LLAMADA	

FIGURE A.1: Evalutation Template



ITEM	SUBITEM	CUMPLE	NO CUMPLE
ACOGIDA			
	Acogida corporativa	• Emplea saludo y presentación corporativa según canal y emisión / recepción • Si en el transcurso de la conversación cambia el interloculor se presenta nuevamente • Continúa con el protocolo de presentación aunque el cliente conteste a su saludo • Comienza la presentación en un tiempo inferior a 3 seg tras recibir la llamada • Está preparado para la siguiente llamada. Sensación positiva. Su presentación es natural, entendible y comercial • El cliente interrumpe la presentación del asesor para exponer directamente el argumento • Por causas ajenas al asesor no se pueda entender la presentación	• No emplea presentación corporativa • Saludo coloquial, falta de concentración. • Si el transcurso de la conversación cambia el interloculor no vuelve a presentarse • Denota cansancio, apatía en el saludo, repetitividad. Su presentación es mécanica, difícil de entender y/o poco comercial.
	Identificación del cliente	• Identifica de forma correcta al cliente o interlocutor. • Pregunta y confirma todos los datos establecidos por protocolo en la política de seguridad • No facilitar información a no tomadores de las pólizas • No ofrece informacion que por motivos de seguridad (LOPD) Generali no desea divulgar	• No identifica al cliente o interlocutor • No pasa pólitica de seguridad • Facilita informacion a no tomador • Ofrece información que por motivos de seguridad Generali no desea divulgar
DESPEDIDA			
	Despedida corporativa	• Antes de finalizar la llamada pregunta al cliente si puede ofrecerle ayuda adicional con fórmulas como ¿puedo ayudarle en algo más? ¿puedo facilitarle alguna información adicional? • Cuando se corta la llamada	• No ofrece ayuda adicional
	Ofrece ayuda adicional	• Utiliza la despedida/s establecida para cada canal y servicio • Su despedida es natural y entendible • Cuando se corta la llamada	• No utiliza la despedida establecida. • Despedida coloquial. • Su despedida es mecánica y difícil de entender. • No se despide, permaneciendo en silencio al finalizar la llamada.

FIGURE A.2: Example of Evaluation Criteria

Bibliography

Azaria, Amos and Tom Mitchell (2023). "The internal state of an llm knows when its lying". In: *arXiv preprint arXiv:2304.13734*.

Chang, Jocelyn J and Edward Y Chang (2023). "SocraHealth: Enhancing Medical Diagnosis and Correcting Historical Records". In: *The 10th International Conf. on Computational Science and Computational Intelligence*.

Chicco, Davide and Giuseppe Jurman (2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC genomics* 21.1, pp. 1–13.

ELIZA, A (n.d.). "I. The early roots of online counseling". In: ().

Gareta (2023). *Can-LLMs-Replace-Human-in-Speech-Analysis*. https://github.com/garetapo/Can-LLMs-Replace-Human-in-Speech-Analysis.

Hawthorn, C. J., K. P. Weber, and R. E. Scholten (July 2023). "BERT-deep CNN: state of the art for sentiment analysis of COVID-19 tweets". In: *Review of Scientific Instruments* 13.99, pp. 1–20. URL: https://link.springer.com/article/10.1007/s13278-023-01102-y.

Park, Youngja and Stephen C Gates (2009). "Towards real-time measurement of customer satisfaction using automatically generated call transcripts". In: *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1387–1396.