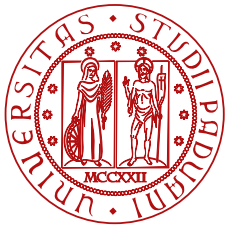

Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Master Thesis in Computer Engineering

Leveraging Large Language Models for Structured Summarization with Evaluation through Reference Matching

Master Candidate

Hilal URUN

Student ID 2048678

Supervisor

Prof. Giorgio Satta

University of Padova

Co-supervisor

Massimo Brunelli

smartKYC

Academic Year
2023/2024

Acknowledgments

I would like to express my deepest gratitude to my academic advisor, Professor Giorgio Satta, for his support, insightful guidance, and encouragement throughout this research project. His expertise was instrumental in helping me overcome the challenges of this thesis and supporting my development as a researcher.

I would like to express my sincere gratitude to my advisors at smartKYC (© 2014-2024 smartKYC Ltd. All rights reserved. Trademark references - UK: UK00003298568, EU: 018827498, US: 7510146), Massimo Brunelli, Noa Goldring, and linguist Aviv Schoenfeld, who worked with me, for providing me with the opportunity to work on this innovative project. Their practical insights and professional guidance contributed significantly to the development and success of this research.

I would also like to thank my family for their unconditional love and support throughout my academic journey. Their constant encouragement and understanding have been a source of strength and motivation to face the challenges of graduate education.

Finally, I would like to acknowledge everyone who contributed to this thesis, whether through insightful discussions, feedback, or support. Your contributions were instrumental in the completion of this work, and I am truly grateful.

To my family
and friends

Abstract

The rapid growth of digital information necessitates effective methods for condensing long technical documents into coherent and concise summaries. This thesis explores the application of structured summarization techniques using large language models (LLMs) to enhance the efficiency and quality of summarizing long texts. The study mainly concentrates upon the advancement of LLMs into building structured summaries and evaluation against human-written summaries based on both qualitative and quantitative metrics.

The work here in this section starts with the selection of a corpus of long documents from various domains and, thus, there is a summary generation task using state-of-the-art LLMs. After this phase, there is an assessment of the comparison between human-crafted and state-of-the-art summaries with respect to coherence, completeness, and correctness. They evaluate a performance of summaries through quantitative performance using statistical measures such as ROUGE, BLEU, and METEOR.

Along with this quantitative assessment, qualitative assessment is also conducted through human evaluation where linguistics experts are asked to evaluate the summaries for readability, relevance, and informativeness. The results from the two assessments provide a comprehensive understanding as to the strengths and weaknesses of LLM-based structured summarization.

The overall outcome of this thesis shall go toward providing further development of summarization technologies and also an understanding of further recommendations that would improve the quality of summaries provided by LLMs. Overall, the study specifies the potential of structured summarization in handling the massive amounts of information generated daily and highlights the upgrades of language models in satisfying the growing demands for information processing.

Sommario

Mentre l'era digitale vede diventare sterminato il corpus delle informazioni, si è reso sempre più urgente il dovere d'inventare metodi che abbiano a costringere la mole di tali informazioni a ristretto contorno. Questo il perché della tesi, che va approfondendo i mezzi collettivi e sistematici, che i grandi modelli linguistici (LLM) offrono, per ottenere sintesi efficace e snella da testi lunghi. Lo studio si concentra, infatti, sopra l'impiego di LLM più alti, e verifica la qualità dei riassunti strutturati che essi producono, colla comparazione fra tali riassunti e i riassunti umani, mediante metriche qualitative e metriche quantitative.

La ricerca comincia col comporre un corpus di scritti lunghi tratti dagli argomenti più vari e dal quale si cavano riassunti strutturati con ricorso a LLM all'avanguardia. Questi riassunti vengono poscia confrontati coi riassunti umani, per stabilire con quale maggior coerenza, massima completezza e minore errore sieno prodotti. A tal scopo s' impiega il quadro della valutazione colla pertinenza, l'obbedienza a certo tipo di coerenza, l'esattezza, misurate attraverso metriche statistiche (ROUGE, BLEU, METEOR).

A ciò s' aggiunge una indagine qualitativa per opera degli uomini addetti al perito giudizio, la quale sotto alla luce della immediatezza, della rilevanza e dell'informatività riflette in qualche modo il risultato. Il confronto dei dati conseguiti nell'una e nell'altra valutazione è destato a far comprendere in che cosa consista il pregio e in che momento la demerito della sintesi strutturata basata sopra LLM.

Tale sommario, vale a dire, se importa un avanzo di conoscenze nel rispettivo campo tecnologico, e se aiuti a comprendere l'andamento degli LLM nella fattiva produzione di riassunti di eccellente fattura; se supplisca altresì ai vuoti lasciati dall'una e dall'altra valutazione per ciò che riguarda la gestione del sempre maggiore volume dePinformazioni e segnatamente per la perpetua esigenza di miglioramento ai modelli linguistici, a norma dei mutati bisogni e degl'indirizzamenti che l'elaborazione delle informazioni prende.

Contents

List of Tables

List of Figures

List of Acronyms

AI Artificial Intelligence

API Application Programming Interface

AWS Amazon Web Services

BART Bidirectional and Auto-Regressive Transformers

BERT Bidirectional Encoder Representations from Transformers

BERTScore BERT Score (an evaluation metric based on BERT embeddings)

BLEU Bilingual Evaluation Understudy

CoT Chain-of-Thought

GenAI Generative Artificial Intelligence

GDPR General Data Protection Regulation

GPU Graphics Processing Unit

GPT Generative Pre-trained Transformer

GPT-4o GPT-4 Optimized

GPT-4 Turbo GPT-4 Turbocharged

GEMINI Refers to a language model developed by Google

JSON JavaScript Object Notation

KPI Key Performance Indicator

KYC Know Your Customer

LLM Large Language Model

LSTM Long Short-Term Memory

METEOR Metric for Evaluation of Translation with Explicit ORdering

Mistral 7B Mistral 7-Billion-parameter Model

NLG Natural Language Generation

NLP Natural Language Processing

RNN Recurrent Neural Network

ROUGE Recall-Oriented Understudy for Gisting Evaluation

smartKYC (Refers to the company smartKYC; not an acronym)

T5 Text-to-Text Transfer Transformer

1 Introduction

Summarizing is the process of reducing the length of a text by extracting the most important information from it, while preserving its core message and overall meaning. Summaries are important tools for information management, allowing individuals to learn about a document without having to read the entire document. Summarization techniques can be divided into two main types: extractive and abstractive. Extractive summarization involves selecting key sentences or phrases directly from the source text, while abstractive summarization produces new sentences that convey the main ideas of the text.

In an era characterized by information overload, the need for effective and efficient summaries is becoming increasingly important. Summaries provide concise and relevant information, allowing for rapid understanding of the text and aiding in decision-making. They are especially important in academic research, news dissemination, legal documents and business reports where large amounts of data must be processed and understood quickly. Summaries help save time and resources by condensing information and facilitate communication by providing better information management.

Summarizing is important in a variety of fields. In academia, researchers may need summaries to quickly review large amounts of literature. In the news industry, summaries help viewers gain information without having to read entire articles. For businesses, summaries help save time by increasing productivity by allowing employees to quickly review key documents. Furthermore, summarizing increases accessibility and allows individuals with varying reading skills to more easily grasp complex information. The ability to summarize effectively is a skill that supports learning, comprehension, and effective distribution of information.

Natural Language Processing (NLP) has become more prominent in the field of summarization, especially with the emergence of Large Language Models (LLMs) such as gpt-3.5-turbo and gpt-4-x. These models leverage deep learning techniques to understand and produce human-like texts, making them powerful tools for summarization. LLMs are trained on large text datasets, which allows them to produce consistent and contextually relevant summaries. They can perform both inferential and abstractive summarization and offer flexible solutions tailored to specific needs. The integration of LLMs into summarization tasks has significantly increased the accuracy, consistency, and fluency of the summaries produced, demonstrating that NLP technologies are evolving day by day.

1.1 Background and Motivation

NLP has made significant progress in recent years, and Generative Artificial Intelligence (GenAI) has emerged as a particularly promising and rapidly developing field. This thesis is being conducted in collaboration with smartKYC, a company that specializes in generating detailed reports on individuals or organizations based on queries. The system used by smartKYC compiles comprehensive documents covering various legal aspects, such as Criminal Collections, Legal Issues, and Arrest Records.

Given the length of the documents produced, it is expected that customers will have difficulty effectively reading and analyzing these extensive documents to make informed decisions. The vast amount of detailed information combined with the complexity of information necessitates a more efficient approach to extract and present the most relevant details, which is addressed by smartKYC. In this context, the need for a concise and informative summary is clear. An effective summary system not only summarizes the essential information, but also provides references to the original documents, allowing customers to access more details when needed. While several open-source models exist for summarization tasks, these models often fall short in specialized applications such as those required by smartKYC. General-purpose models that perform a variety of tasks such as summarization and translation often lack the specificity and precision required for this particular application. To address this gap, various strategies have been considered, including instruction tuning, fine-tuning, and prompt engineering. Given the hardware and time constraints, prompt engineering has been determined to be the most practical and efficient solution.

After extensive research and testing of multiple models including gpt-3.5-turbo, gpt-4-turbo-preview, gpt-4-turbo, gpt-4o, gemini-1.5-pro, gemini-1.0-pro

and LLama3, gpt-4-turbo-preview was currently selected for the test stage due to its superior performance in this context. However, tests and comparisons between gpt-4o and gpt-4-turbo-preview are still ongoing. Specific prompts were developed for each crime type to ensure relevant information extraction. Through iterative improvements and testing, the performance of the system was evaluated using human-written summaries as reference points along with the creation of relevant evaluation metrics. In particular, this approach provides a novel contribution to the field due to the limited research on performance evaluation using gpt-4o and gpt-4-turbo-preview for structured summarization tasks. Additionally, the system was further refined by generating summaries of summaries for each page with continuous improvements based on performance metrics. The overall goal is to develop an automated summarization and evaluation system specifically tailored to the needs of smartKYC, thereby increasing the efficiency and accuracy of information provided to clients.

The motivation for this research stems from the intersection of broader trends in NLP and the specific needs of smartKYC. Rapid advances in GenAI provide an opportunity to significantly improve automated summarization systems. By leveraging state-of-the-art models such as the gpt-4-x family, this project aims to address a critical challenge facing smartKYC clients: the ability to quickly and accurately understand extensive and complex legal documents.

One reason for this is that traditional methods of reading and analyzing detailed reports are not only time-consuming but also prone to human error. This research aims to improve decision-making processes to develop a robust summarization system, providing clients with clear, concise, and relevant information. This approach not only saves time, but also ensures that important details are not overlooked.

In summary, this thesis stems from the goals of advancing NLP technologies and addressing the specific needs of the smartKYC platform. The development of an effective summarization and evaluation system aims to make a meaningful contribution to both the academic community and practical applications in the field of legal document analysis.

1.2 Problem Statement

In the broader context of NLP, summarizing large and complex documents is a significant challenge. As data volumes continue to grow, the ability to extract important information from extensive text is becoming increasingly critical across a variety of domains, including legal, financial, and compliance services.

Existing summarization models often struggle to effectively handle the length, complexity, and domain-specific nuances of such documents, resulting in summaries that are either too superficial or lack essential detail.

In this context, smartKYC, a company that provides comprehensive reports on individuals and organizations, faced a challenge as they wanted to summarize their large and complex legal documents. These documents, which cover aspects such as Crime Totals, Legal Issues, and Arrest Records, are important for clients who need to make informed decisions. However, the abundance and detail of information presented can make it difficult for decision-making and clients to efficiently extract the relevant insights they need.

Existing pretrained summarization models are largely general purpose and it is desired to test whether they meet the specific needs of the smartKYC domain. We want to study whether they are capable of producing concise, accurate and informative summaries that capture critical aspects of these legal documents and also provide references to the original sources.

This thesis addresses the sub-problem of developing a summarization system for smartKYC that can effectively process and summarize these complex legal documents. The system should produce summaries that are not only concise and informative, but also linked to relevant sections of the original documents, providing easy access to detailed information when needed.

In addition, a robust evaluation framework is needed to evaluate the quality and effectiveness of the summaries produced, especially in the context of structured, domain-specific content. The current lack of standardized evaluation methods for models such as gpt-4-x in such specialized tasks presents an additional challenge that this thesis aims to overcome.

In summary, the overall problem involves the difficulties of summarizing large documents in general, with a specific focus on addressing the summarization needs in the domain of smartKYC. The aim is to develop a specialized solution that not only meets these specific needs, but also constitutes a reliable method for evaluating the performance of such a system.

1.3 Objectives

The main objective of this thesis is to develop a robust and efficient summarization system that addresses the challenges of summarizing large and complex legal documents in the smartKYC domain. To achieve this overall goal, the thesis is structured around the following specific objectives:

- Conducting a comprehensive review of existing summarization models

and techniques, focusing specifically on their applicability to large, complex, and domain-specific documents.

- Identifying the limitations of general-purpose LLMs in handling smartKYC legal documents.
- Evaluate various state-of-the-art GenAI models on their performance in summarizing legal documents
- Selecting the most appropriate model based on criteria such as accuracy, relevance, and computational efficiency
- Developing specific prompts for each crime type (e.g., Criminal Collections, Legal Issues, Arrest Records) to ensure that relevant information is extracted with precision.
- Ensuring that summaries generated include references to original sections of the documents and ensure that clients can access detailed information when needed.
- Developing a comprehensive evaluation framework to evaluate the quality and effectiveness of the summaries generated.
- Using human-written summaries as reference points and determine relevant evaluation criteria to measure system performance.
- Conducting systematic performance evaluations to identify areas for improvement and improve the summarization system accordingly.
- Implementing iterative improvements to prompts and the summarization system based on the evaluation results.
- Exploring methods to automate the summarization and evaluation processes, increasing the overall efficiency and accuracy of the system.
- Providing insights and recommendations for future research and practical applications in the field of legal document analysis and summarization.

This thesis aims to develop a specialized hashing system to meet the needs of the smartKYC platform by achieving these goals, while also adding valuable insights to the broader fields of NLP and GenAI.

1.4 Contributions

This thesis makes significant contributions to both the academic community and practical applications in the business sector.

The research conducted in this thesis provides a detailed comparison of the performance of various LLMs in the task of summarizing complex legal documents. By evaluating models such as the Open AI family of LLMs in the context

of structured summarization, this study fills a gap in the existing literature where little research has been conducted on implementing and evaluating LLMs for domain-specific summarization tasks. Additionally, the thesis aims to provide insights into how these models can be adapted and optimized to meet specific needs in specialized domains by investigating the potential of fine-tuning pre-trained models for task-specific purposes. The findings from this research are expected to contribute to the ongoing development of NLP techniques and models, providing a foundation for future work in both summarization and legal document processing.

From a business perspective, the thesis provides a practical solution that improves the efficiency and effectiveness of smartKYC's services. By developing a summarization system that provides faster and more accurate insights into the legal status of questioned individuals or organizations, the project directly benefits the company's customers. The ability to quickly understand and evaluate documents allows customers to make more informed decisions, save time, and reduce the risk of missing critical details. This contribution demonstrates the practical value of integrating advanced NLP technologies into business processes, highlighting the potential of AI-driven solutions to improve service delivery in the legal and compliance sectors.

In summary, this thesis aims not only to advance academic knowledge in the field of NLP and summarization, but also to provide tangible benefits to business applications in increasing the efficiency and accuracy of large document analysis.

2

Literature Review

2.1 Overview of Large Language Models

LLMs are powerful AI systems trained on large amounts of text data. These models can understand, produce, and process human language in ways that often mimic human capabilities. They are also built on transformative architecture that allows them to capture long-range dependencies in text. Popular examples include GPT models and Bert. LLMs have a wide range of applications, from natural language processing tasks to creative content generation and customer service.

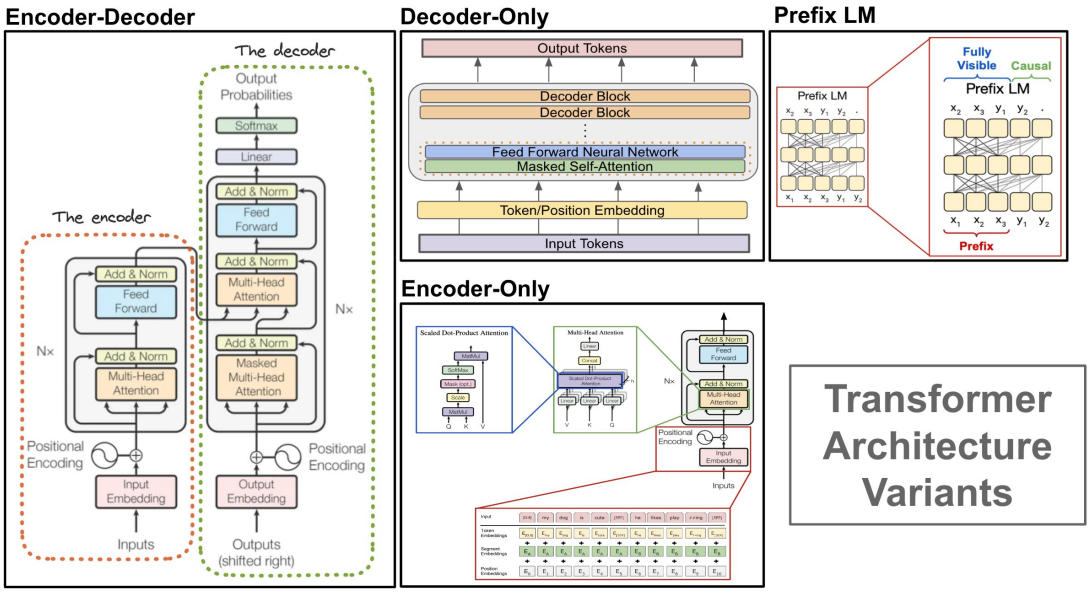

Figure 2.1: Transformer Architecture

The Transformer architecture, introduced in the paper "Attention Is All You

Need" [67], has revolutionized the field of natural language processing. Unlike recurrent neural networks (RNNs) and convolutional neural networks (CNNs), which process sequences sequentially, Transformer relies solely on attention mechanisms to capture dependencies between different locations in the input sequence.

The main components of the Transformer architecture are the Encoder, Decoder and Attention Mechanism. Each of the encoders consists of multiple identical layers. Each layer contains a self-attention mechanism that calculates the weighted sum of all input tokens to determine the importance of the token relative to the current token. It contains feed-forward neural networks that apply a non-linear transformation to the output of the self-attention mechanism. The decoder is similar to the encoder, but has an additional masked self-attention mechanism to prevent the model from paying attention to future tokens during training. It also contains an encoder-decoder attention mechanism to capture the dependencies between the input sequence and the output sequence. The Attention Mechanism can be considered as the core of the Transformer architecture. It calculates the weighted sum of the input tokens according to their relevance to the current token. It uses a query, key and value mechanism to calculate the attention weights.

Advantages of the transformer architecture are, unlike RNNs, transformers can process the entire input sequence in parallel, making them more efficient for long sequences. Also its attention mechanism allows the model to capture dependencies between distant parts of the input sequence. In addition, transformers have achieved state-of-the-art results on various NLP tasks, including machine translation, text summarization, and question answering.

2.1.1 BART (Bidirectional Encoder Representations from Transformers)

BART, introduced in the paper "BART: A Bidirectional Encoder Representations from Transformers for Sequence-to-Sequence Prediction" (Liu et al., 2019),is a transformer-based model that functions as a denoising auto-encoder for sequence-to-sequence tasks. The architecture of BART combines the strengths of both BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer). It is pre-trained by corrupting text with an arbitrary noising function and then learning to reconstruct the original text. This training process makes BART particularly effective for a variety of tasks, including text generation, machine translation, and especially summa-

rization.

The model consists of an encoder-decoder structure: the encoder processes the input text bidirectionally and captures the context from both directions, while the decoder produces the output text autoregressively and predicts each token separately. This structure allows BART to address tasks where both understanding the full context and producing consistent outputs are important.

Key differences between BART and the original Transformer:

- Bidirectional encoder: BART uses a bidirectional encoder, which allows it to process the input sequence in both directions, capturing context from both the left and right.
- Noisier data: BART is trained on a noisier dataset, which includes random token deletions and permutations. This helps the model to become more robust to noise and variations in the input data.
- Sequence-to-sequence prediction: BART is specifically designed for sequence-to-sequence tasks, making it well-suited for tasks like summarization.

BART has been widely adopted for summarization tasks, particularly for abstractive summarization, where the goal is to generate new sentences that convey the key points of the original text. In its initial implementation, BART was shown to outperform several other models on benchmark summarization datasets like CNN/Daily Mail and XSum, achieving state-of-the-art results. One of the primary challenges with BART, as noted in several studies, is maintaining factual consistency in the generated summaries. While the model is capable of producing fluent and coherent summaries, it sometimes introduces inaccuracies or hallucinations, where information not present in the original text is generated. Also its performance can degrade when dealing with long documents, as the model's input size is limited by the transformer architecture. Summarizing long texts often requires either truncating the input or breaking it into smaller chunks, which can lead to loss of important context or coherence issues in the summary.While BART excels in generating fluent text, it is less suited for structured summarization tasks that require generating summaries with specific formatting or reference linking back to the original document sections. This limitation is particularly evident in fields like legal and technical document summarization, where the structure of the output is important.

2.1.2 Llama2 and Llama3

Llama2 and Llama3 are advanced LLMs, part of the Llama family designed to push the boundaries of natural language understanding and generation. Developed by Meta AI (formerly Facebook AI), these models are versions of the

original LLaMA model. Llama2 and Llama3 are built on a similar transformative architecture to other LLMs such as BERT and gpt, but with enhancements that allow them to handle longer contexts more effectively and produce more detailed outputs.

Llama2 was introduced as a mid-range model focused on efficiency, aiming to provide high performance without requiring extensive computational resources. It was specifically designed to meet the needs of tasks that require understanding long documents, making it suitable for applications where capturing context from extended text, such as summarization, is crucial.

Llama3 builds on the capabilities of Llama2, offering improvements in model architecture and training data. It is designed to handle even more complex language tasks, including cross-language applications and domain-specific text generation. Llama3's architecture incorporates improvements in attention mechanisms and memory management, allowing it to maintain consistency across longer sequences and produce more contextually relevant summaries.

Despite the advancements offered by Llama2 and Llama3, several challenges remain, particularly in the context of summarization tasks.One of the challenges noted in several studies is maintaining contextual relevance when summarizing very long documents. While Llama2 and Llama3 can handle extended text sequences better than their predecessors, there are still issues with ensuring that the summary accurately reflects the most important aspects of the entire document.The other challenge is similar to other LLMs, Llama2 and Llama3 can sometimes generate information that is not present in the original text, a phenomenon known as "hallucination." This issue has been highlighted in summarization tasks where the models occasionally introduce errors or fabricate details, leading to factual inaccuracies.

2.1.3 GPT-3.5-TURBO

Developed by OpenAI, gpt-3.5-x is a major iteration in the GPT series. Building on the success of its predecessor gpt-3, which was known for its ability to produce consistent and contextually relevant text across a wide range of applications, gpt-3.5 introduces improvements to both the model architecture and training methodologies.

The model consists of a 175 billion parameter transform architecture, making it one of the largest and most powerful language models available at the time of its release. The model is designed to perform a variety of NLP tasks, including text generation, translation, question-answering, and summarization. A large

dataset of diverse internet texts was used to pre-train the model, enabling it to capture a broad understanding of language patterns, context, and semantics.

The improvements in gpt-3.5 are not only in scale, but also include improvements to the training process, such as better handling of long-range dependencies and improved contextual understanding. These improvements suggest that gpt-3.5 could be effective for tasks that require synthesizing information from long or complex texts, such as automatic summarization. The model can demonstrate remarkable capabilities in automatic summarization, both inferential and abstractive formats. Studies have shown that it is capable of producing concise, coherent, and contextually relevant summaries for a wide variety of texts, from news articles to technical documents. According to the study, gpt-3.5 outperformed its predecessors and several other models in producing summaries that closely match human-written summaries in terms of informativeness and fluidity.

One of the main challenges in summarization is dealing with long documents where the model must understand and distill large amounts of information. Despite its strengths, gpt-3.5 also faces several challenges in summarization. One recurring problem is that summaries are frequently produced that contain inaccuracies or "hallucinatory" details that are not present in the source text. This can be problematic in contexts where accuracy is critical, such as summarizing legal documents or scientific papers. Like other major language models, gpt-3.5 can sometimes reflect biases present in the training data, leading to skewed summaries or failure to capture a balanced perspective.

They have demonstrated solid performance on both inferential and abstractive summarization tasks. Studies show that they consistently outperform their predecessors, including Bard and gpt-3.5-turbo, in producing coherent and concise summaries that accurately reflect the content of the source material. This helps establish the models as a powerful tool for summarizing long documents such as research papers, legal texts, and technical reports. Studies have noted that their cross-domain performance is significantly superior to previous models, which is attributed to their improved contextual understanding and integration of external knowledge. This allows the models to produce summaries that are not only accurate, but also tailored to the specific nuances of each domain.

2.1.4 GEMINI-1.0-PRO AND GEMINI-1.5-PRO

The Gemini models developed by Google represent a significant advancement in the field of large language models in NLP. Designed as successors to

the Bard model, the models push the boundaries of what LLMs can achieve in understanding, generating, and summarizing complex text across a variety of domains. Building on Google's extensive research on Transformer architectures, the model includes specific optimizations that increase its ability to handle nuanced language tasks such as summarization, translation, and question-answering.

The model's architecture features several innovations aimed at improving both efficiency and accuracy, particularly when processing large volumes of text while preserving context in long content. A key feature is the model's ability to dynamically integrate external information sources, allowing it to produce more informed and contextually relevant summaries. This ability makes it particularly well-suited for applications that require the synthesis of information from multiple documents or databases.

The models are also capable of summarizing long documents, a critical capability given the increasing volume of extensive textual data across industries. They can condense long documents while preserving the essential meaning and intent of the original text. This can be achieved through an advanced memory management system that preserves context across long text sequences and reduces the risk of skipping important information during the summarization process. However, despite their strengths, the models face several challenges. In terms of contextual consistency, the model generally achieves a high level of consistency, but sometimes struggles with documents that are particularly long or complex. Cases have been observed where their summaries lose focus, especially when the source material contains multiple equally important points. Another concern is bias in summarization. Like many large language models, Gemini models tend to reflect inherent biases in the training data, especially when summarizing content that includes subjective perspectives or controversial topics. This can sometimes result in summaries that are skewed interpretations of the source material.

Until 2024, related studies have examined the application of Gemini in various domains. In cross-domain summarization, studies have found it to be particularly effective at producing summaries that accurately capture the essence of text across domains, outperforming models such as gpt-4 and BERT. However, this effectiveness has been lost with the introduction of gpt-4o and later models.

In conclusion, gemini models represent a significant advance in large language models, offering robust capabilities for automatic summarization across domains. Its strengths in cross-domain and long-document summarization, combined with its ability to integrate external information, make it a valuable

tool for both academic and commercial applications. However, challenges related to contextual consistency, bias, and computational demands remain areas that require further research and development.

2.1.5 GPT-4o

gpt-4o is an advanced version of the gpt-4 series designed to meet the increasing demand for more efficient and accurate language models in NLP tasks. The "o" in GPT-4o refers to improvements made to the model to improve its performance, especially in tasks such as summarization, where understanding context and producing concise output are critical.

The model is a versatile model developed by Open AI that can process text, audio and visual data and produce it in real time. This model is considered a significant milestone in the field of artificial intelligence and offers many innovations compared to previous versions. It is also a language model trained on a huge amount of data. During this training process, the model learns the structure, meaning and context of the language. The basic working principle of the model is to predict the possible next element on a given input (text, audio or image). These predictions are performed by a complex neural network with billions of parameters. One of the most important features of the model is that it is multi-modal and the output is repeatable. In other words, it can understand both text and visual data and switch between them. In this way, it can perform tasks such as answering a visual question or visually summarizing a text.

One of the most striking features of the model is its ability to summarize texts very effectively. The model can turn long and complex texts into short and understandable summaries. These summaries preserve the main ideas of the original text while omitting unnecessary details.

As a result, the publication of the model is considered an important step in the field of artificial intelligence. The model's versatile structure and strong summarizing ability allow it to be used in many areas. Especially in today's world where access to information has become easier, the ability of models such as gpt-4-o to summarize long texts quickly and accurately is of great importance.

2.2 Summarization Techniques

Summarization techniques are of critical importance in the field of NLP, especially as the volume of textual data continues to grow exponentially. The need for concise, informative, and contextually relevant summaries has led to

the development of a variety of summarization methods. These techniques can generally be divided into two main types: extractive and abstractive summarization. Additionally, structured summarization has emerged as a specialized approach that combines the advantages of both methods while addressing specific challenges in structured data contexts.

2.2.1 Extractive and Abstractive Summarization

Extractive summarization involves extracting sentences, phrases, or sections directly from the source text to create a summary. This technique relies on identifying the most critical sections that best represent the overall content of the document. The primary advantage of extractive summarization is its simplicity and the ability to produce grammatically correct summaries because it uses sentences directly from the original text. However, extractive methods often struggle to capture the nuanced meaning of the text, and the resulting summaries may lack coherence or natural flow.

A variety of algorithms and models have been developed to improve extractive summarization, including graph-based methods such as Text-Rank and machine learning approaches that use features such as sentence position, word frequency, and semantic similarity. Recent developments have included transformative models such as BERT to improve the extraction process by better understanding the context and importance of sentences in a document. Despite these advances, extractive summarization is inherently limited because it relies on the original text, making it less flexible than abstractive techniques.

Abstractive summarization, on the other hand, involves creating new sentences that capture the essence of the original text. This approach is more complex because it requires the model to understand the content and then rephrase or condense it into a summary. Abstractive methods aim to produce more coherent and natural summaries than those generated by inferential methods.

Recent developments in abstractive summarization have been driven by advances in neural network architectures, particularly the introduction of transformer-based models such as gpt, bart, and t5. These models can produce summaries that are not only concise but also contextually rich and semantically accurate. Abstractive summarization is most often used in scenarios where the summary must convey the original meaning in a more human-like, readable form. However, the complexity of this task does not change the fact that abstractive models sometimes produce inaccurate or overly general summaries; this is a challenge that ongoing research continues to address.

2.2.2 Structured Summarization

Structured summarization is a technique that aims to produce summaries that not only condense information but also organize it in a way that conforms to specific formats or patterns. This approach can be particularly useful in areas where the summary must follow a structured format, such as legal, medical, or financial documents. Structured summarization often integrates elements of both inferential and abstractive summarization, while also ensuring that the summary conforms to predefined structures. Structured summarization goes beyond summarizing text; it aims to organize content in a way that is meaningful to the end user. One of the key challenges in structured summarization is ensuring that the content generated is both accurate and consistent with the desired output structure. This often requires the integration of domain-specific knowledge and the ability to produce summaries that are contextually aware of the organization of the document and the relationships between different sections. Techniques such as template-based summarization, where the model is driven by predefined templates, and the use of large language models such as gpt-4o that can understand and produce structured content are increasingly being used to address these challenges.

In summary, structured summarization represents a significant step forward in the field of automatic summarization, offering the potential to generate summaries that are not only concise and informative but also organized in a way that enhances their usability. As LLMs continue to evolve, their ability to generate structured summaries will likely improve, making this a promising area for future research and application.

2.3 Evaluation Metrics for Summarization

Evaluating the quality of automatic text summarization is one of the critical aspects of NLP research and application. The effectiveness of summarization models is often evaluated using both automatic and human evaluation methods. Automatic evaluation metrics provide a quantitative measure of the similarity between machine-generated summaries and human-written reference summaries. This section examines the most commonly used evaluation metrics for summarization, including BLEU, METEOR, ROUGE, BERTScore, and GEval, and the role of human evaluation.

2.3.1 BLEU (Bilingual Evaluation Understudy

BLEU is a precision-based metric originally developed to evaluate machine translation but adapted for summarization tasks. It measures the overlap between n-grams of the candidate summary and the reference summary. BLEU calculates a modified precision score for n-grams, typically up to four grams (BLEU-4), and includes a brevity penalty to penalize overly short candidate summaries.

Considering its advantages, it is easy to calculate and interpret. It also has no language restrictions, which is why it is a widely used evaluation metric. On the other hand, BLEU emphasizes precision over recall and may ignore missing but important content in the candidate summary. However, it does not take into account semantic equivalence through synonyms or paraphrases, and it does not consider the order of sentences or the overall coherence of the summary. Despite its limitations, BLEU remains a common metric for summarization due to its simplicity. However, it is often used in conjunction with other metrics that take into account recall and semantic similarity.

Formula :

$$\text{BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log P_n\right) \tag{2.1}$$

where:

- BP is the brevity penalty.
- P_n is the precision for n-grams of length n.
- w_n is the weight assigned to the n-gram of length n (usually uniform, e.g., 1/4 for four-grams).

2.3.2 METEOR (Metric for Evaluation of Translation with Explicit ORdering)

METEOR is a metric that was introduced to address some of the shortcomings of BLEU by combining both precision and recall with synonym and stemming. It calculates the harmonic mean of single-gram precision and recall, and recall is weighted higher than precision. METEOR also includes a penalty score for word order differences.

It provides a more holistic assessment by considering both aspects in the calculation. It also takes into account synonyms and morphological variations through stemming and synonym matching.It generally shows better agreement with human evaluations than BLEU. On the other hand, it requires more complex computation than BLEU and language-specific resources such as stemmers and synonym databases.

METEOR's consideration of semantic similarity makes it suitable for summarization assessment. The ability to capture paraphrasing and synonyms allows for a more flexible assessment of summary quality. Formula :

$$\text{METEOR} = \left(\frac{10 \cdot \text{Precision} \cdot \text{Recall}}{\text{Recall} + 9 \cdot \text{Precision}}\right) \times (1 - \text{Fragmentation Penalty}) \tag{2.2}$$

where:

- Precision is the proportion of words in the candidate summary that are also in the reference summary.
- Recall is the proportion of words in the reference summary that are also in the candidate summary.
- Fragmentation Penalty reduces the score based on the disjointedness of matched words in the candidate summary.

2.3.3 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE is a set of metrics specifically designed for summarization assessment. It focuses on recall by measuring the overlap of n-grams, word sequences, and word pairs between the candidate summary and the reference summaries. ROUGE emphasizes the coverage of important content, and multiple variants allow for different aspects of assessment. However, it may reward longer summaries that contain more content but are less concise, and it does not take into account the readability or logical flow of the summary.

Variants of ROUGE:

- ROUGE-N: Measures n-gram recall between candidate and reference summaries. Commonly used with n=1 (ROUGE-1) and n=2 (ROUGE-2).
- ROUGE-L: Uses the longest common subsequence to assess the reference coverage of the summary.

- ROUGE-S: Considers skip-bigram associations.

ROUGE is considered the standard metric for summarization assessment because of its focus on content coverage. It is particularly useful for comparing systems in terms of how much important information they capture from the source text. The way it actually works is as follows:

- The most common variant is ROUGE-N, where "N" represents the length of the n-grams (e.g., unigrams, bigrams). ROUGE-N measures the overlap of n-grams between the candidate summary and the reference summary.

$$\text{ROUGE-N} = \frac{\sum_{\text{gram} \in \text{Reference}} \text{Count}_{\text{match}}(\text{gram})}{\sum_{\text{gram} \in \text{Reference}} \text{Count}(\text{gram})} \tag{2.3}$$

  where:

  - $\text{Count}_{\text{match}}(\text{gram})$ is the number of n-grams in the candidate summary that also appear in the reference summary.
  - $\text{Count}(\text{gram})$ is the total number of n-grams in the reference summary.

- ROUGE-L measures the longest common subsequence (LCS) between the candidate and reference summaries. LCS is a sequence that can appear in both summaries in the same order without necessarily being contiguous. ROUGE-L thus captures the overall structural similarity between the summaries.

- ROUGE-S (Skip-Bigram Co-occurrence) measures the overlap of skip-bigrams, which are pairs of words that occur in the same order in both summaries but with any number of intervening words. This allows ROUGE-S to account for more flexible word orders.

- ROUGE-W (Weighted LCS) is a variant of ROUGE-L that applies a weighting to the LCS, giving more importance to longer subsequences.

2.3.4 BERTScore

BERTScore is a newer metric that evaluates summaries based on semantic similarity using contextual embeddings from pre-trained models like BERT. It calculates similarity scores between tokens in candidate and reference summaries using the cosine similarity of their embeddings.

It captures semantic similarities beyond exact word matches, leading to more accurate evaluations by taking into account the context in which words appear. It shows higher correlation with human evaluations than traditional metrics.However, it requires significant computational resources due to the use of large pre-trained models, and the interpretation of scores can be less intuitive than traditional metrics.

BERTScore addresses many of the limitations of n-gram-based metrics by evaluating summaries at the semantic level. It is particularly useful for evaluating the quality of abstract summaries that may use different wording than reference summaries.

Formula :

$$\text{Precision} = \frac{1}{|C|} \sum_{x \in C} \max_{y \in R} \cos(x, y) \tag{2.4}$$

$$\text{Recall} = \frac{1}{|R|} \sum_{y \in R} \max_{x \in C} \cos(x, y) \tag{2.5}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.6}$$

where:

- C is the set of words in the candidate summary.
- R is the set of words in the reference summary.
- $\cos(x, y)$ represents the cosine similarity between the embeddings of word x in the candidate summary and word y in the reference summary.

2.3.5 GEval (Generalized Evaluation Metric)

GEval is a generalized assessment framework that integrates multiple metrics and adapts to various NLP tasks, including summarization. It aims to provide a more comprehensive assessment by combining precision, recall, and semantic similarity aspects. It incorporates multiple assessment aspects into a single framework and can be adapted to different NLP tasks and specific assessment needs. It also allows for the weighting of different components according to their importance.However, the generality of the framework can lead to complexity in implementation and interpretation. It has not been adopted as widely as other metrics, making comparisons with other studies more difficult.

GEval's flexibility makes it suitable for complex assessment scenarios where multiple quality aspects need to be considered. It can be customized to prioritize specific assessment criteria related to a specific summarization task.

- GEval evaluates text based on several key dimensions, which may include:
- Fluency: The grammatical correctness and naturalness of the text.
- Coherence: The logical flow and structure of the text.
- Relevance: The degree to which the generated text covers the key content of the reference text.
- Consistency: The internal logical consistency of the generated text.
- Diversity: The variety in word choice and sentence structure.
- Each dimension is evaluated separately, often using specific sub-metrics tailored to that dimension. The results are then aggregated using a weighted sum to produce a final GEval score. The weights can be adjusted depending on the importance of each dimension for the specific task.
- GEval can incorporate both reference-free (e.g., fluency, coherence) and reference-based (e.g., relevance, ROUGE-based similarity) evaluations, making it versatile for different applications.
- GEval often leverages pre-trained models, like BERT or GPT, to assess fluency, coherence, and relevance. These models help capture the deep contextual and semantic understanding necessary for a comprehensive evaluation.

$$\text{GEval} = \sum_{i=1}^{k} w_i \cdot S_i \tag{2.7}$$

where:

  - S_i is the score for the ith evaluation dimension (e.g., fluency, relevance, coherence).
  - w_i is the weight assigned to the ith dimension.
  - k is the total number of evaluation dimensions.

2.3.6 Human Evaluation

While automated measures provide valuable quantitative assessments, human evaluation is also necessary to gain a comprehensive understanding of abstract quality. Human raters can assess aspects that are difficult to capture with automated measures, such as coherence, readability, factual accuracy, and overall usefulness. However, human evaluation is intensive and time-consuming. Different raters may have different opinions, which can lead to inconsistent evaluations. However, it is impractical to evaluate large numbers of abstracts. Human evaluation is important for tasks where nuanced judgment is required. It is often used in conjunction with automated measures to provide a balanced evaluation of abstracting systems.

- **Subjective Assessment**: Human evaluators assess various qualitative aspects of the text, including fluency, coherence, relevance, readability, and accuracy.
- **Evaluation Criteria**:
  - Fluency: How grammatically correct and natural the text sounds.
  - Coherence: The logical flow and consistency of ideas in the text.
  - Relevance: The degree to which the text covers the key points or content intended by the summary or translation.
  - Readability: How easy it is to read and understand the text.
  - Accuracy: For translation tasks, the correctness of the translated content in relation to the source.
- **Scoring Methods**:
  - Likert Scale: Evaluators rate the text on a scale (e.g., 1 to 5) for each criterion.
  - Binary Judgments: Evaluators make pass/fail judgments based on specific criteria.
  - Ranking: Multiple outputs are ranked in order of preference or quality.
- **Inter-Annotator Agreement**: To ensure reliability, multiple human evaluators are used, and their judgments are compared using measures like Cohen's Kappa or Krippendorff's Alpha.

Evaluation metrics are crucial for the development and benchmarking of summarization systems. Traditional metrics such as BLEU, METEOR, and ROUGE have been widely used but have limitations in capturing the semantic similarity and qualitative aspects of summaries. Recent developments such as BERTScore address some of these limitations by using contextual embeddings. GEval provides a flexible framework for comprehensive evaluation, while human evaluation provides an irreplaceable qualitative evaluation despite its resource requirements.

In practice, combining multiple automated metrics with human evaluation provides a broader picture of a summarization system's performance. This multifaceted approach allows researchers and practitioners to understand the strengths and weaknesses of their models, guiding further improvements.

3 Methodology

3.1 Description of The System in smartKYC

smartKYC is working on automation of Know Your Customer (KYC) and background check processes. Its technology drives faster, better and more cost-effective KYC at every stage of the relationship, freeing up human effort to focus on decision-making instead of laborious research. It also combines AI with linguistic and cultural sensitivity and deep domain knowledge to set new standards for KYC quality, transform productivity and ensure compliance compliance.

At the heart of the smartKYC system is a search capability that seamlessly connects to a wide range of data sources, both structured and unstructured. These sources span public databases, proprietary registries, professional subscriptions and internal blacklists, enabling a more comprehensive KYC process. By integrating data from multiple databases, it gathers information from open web media archives, corporate director and shareholder databases, legal decisions, biographical data, national corporate registries and more. This provides a comprehensive review of the watched entity or institution from which the information was collected, while also helping to ensure that critical information is not overlooked.

It also provides a search and analytics platform that provides risk-related and contextual multilingual options for third parties. This feature is important in today's business world where information can be available in multiple languages. By offering multilingual support, smartKYC enables customers to access and understand relevant information regardless of the language it is presented in, thus expanding the platform's applicability and effectiveness across geographies.

The platform helps in both identifying and analyzing potential customers, customer due diligence, and gaining insights on a person or organization basis in various situations, including periodic updates. In lead management and customer on-boarding, smartKYC helps identify and evaluate potential customers or business partners. The system helps automate the process of updating and verifying customer information for periodic reviews and mass remediation projects, thereby accounting for ongoing compliance or risks. It provides detailed analytics and reporting for high-risk customers or organizations that require additional due diligence. Customers can search for individuals or companies to gather comprehensive information and access this information in various formats. Users can view all documents with key information highlighted for quick reference of important details. Additionally, the platform provides partial information extracted from documents categorized and classified by specific topics such as Crime, Criminal Records, Biography, and more. Users can also download files for offline review and record keeping of the collected information and read information by classification, enabling efficient navigation between topics of interest.

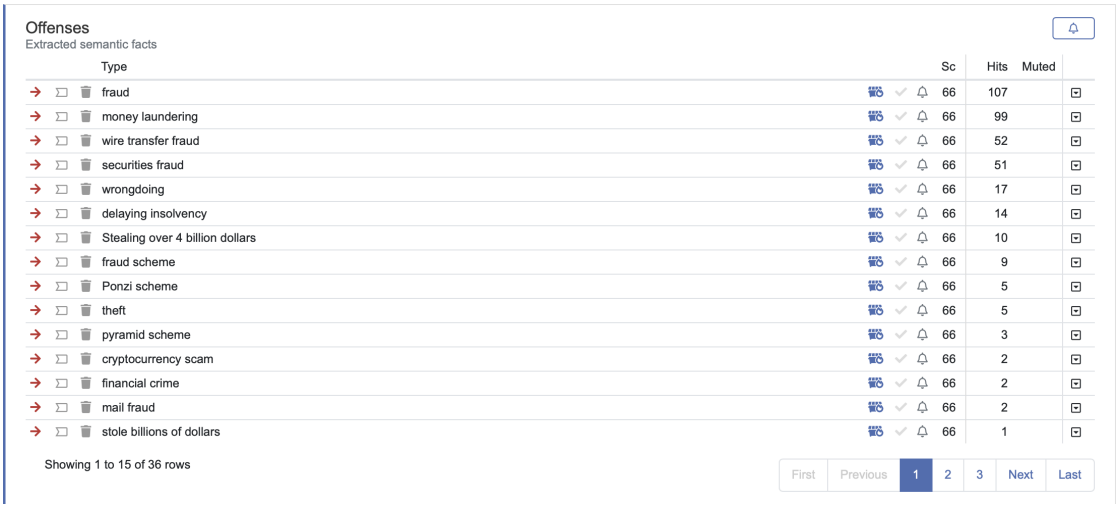

Figure 3.1: smmartKYC platform Offense Aggregation Initial format

When the system is examined, the large amount of information collected poses a challenge for users to review all of the data. In order to make a decision, all documents marked as risky had to be read and reviewed. This challenge revealed the need for an effective summary structure to improve the user experience and shorten and facilitate the decision-making process. In order to meet this need, the project carried out in partnership with smartKYC aimed to develop a structure that could produce informative summaries based on the themes determined first in line with the information collected in the company's database,

and then based on the dimension cluster where the themes were segmented.

Before implementing this structure in smartKYC, a development environment hosted on company servers was created to create a prototype. First of all, the aim of the summary was to improve the user experience by highlighting the important information in the content and obtaining easy-to-read user-friendly summaries, and thus reviewing the referenced reference documents instead of reviewing all the collected documents. This would also provide an improvement in the decision-making and insight-getting process about the entity. In addition, the structure allows traceability to be maintained thanks to the references added to the summaries and allows users to access the entire content by going to the linked document to verify certain sections when necessary. In addition, the summarization process has been carried out in three consecutive stages as manual, semi-automatic and autonomous. Although autonomous feature has not been achieved yet, work continues in that direction.

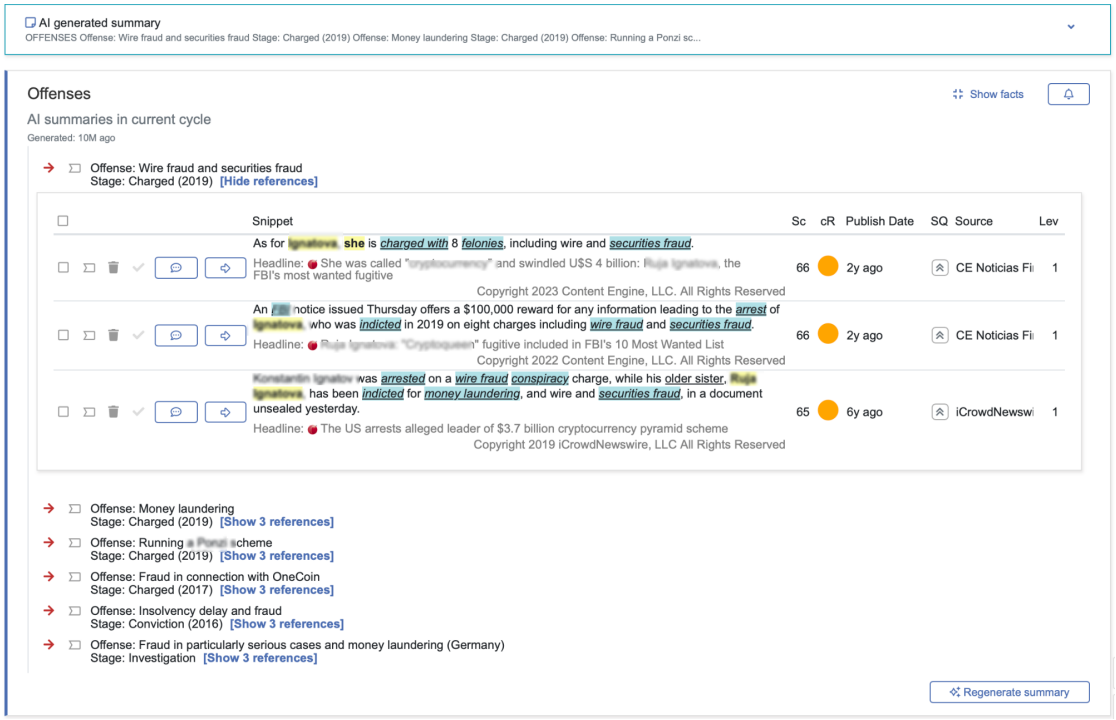

Figure 3.2: smmartKYC platform Offense Aggregation Final format

In summary, smartKYC exemplifies a comprehensive approach to automate KYC and background check processes through AI and NLP technologies. By integrating various data sources and providing multilingual capabilities, smartKYC provides a robust platform for due diligence and compliance. Implementing a structured summarization system addresses the need for efficient information processing, enhancing user experience, and supporting informed decision making. This methodological approach forms the basis for examining

the optimal approach using large language models to produce structured and contextually relevant summaries, which is the primary focus of this thesis.

3.2 Data Collection and Preparation

This section provides information about data collection and preparation of collected data, which is an important step for the summary structure to be developed for the smartKYC platform. Factors such as the appropriateness of the data used in the context, the correct information, and the reliability of the data source are important in order for the large language model to reason and make correct inferences in line with the characteristics determined for the entity.

In the data collection process, the first thing to do is to decide on entities with different nationalities and multiple language options (person or company) and use smartKYC's query feature to collect information about the entities. The platform offers both single search and bulk search options. While single search allows you to send a query to one person or organization at a time, bulk search facilitates multiple searches at once. During these searches, users can enter detailed information to narrow down the results, including demographic data of the subject, preferred languages for source materials, name variants, and specific database providers and search engine choices. This level of detail increases the accuracy and relevance of the search results.

Upon completion of a search, the system generates a review page that initially includes labeled information such as risk levels and source credibility scores. The content in each review is organized into primary categories known as "dimensions," which are then subdivided into "frames." For example, the "Sales Intelligence" dimension includes the "Lifestyle" frame. The structure, segmented by themes and topics, helps classify information. Since Large Language Models have a context window/length restriction, and since multiple topics can be included in a context window at the same time, disadvantages such as hallucination and lack of information can be encountered, the summarization application was designed according to the smallest theme frames. In the next stage, it was decided to summarize the frames included in the same dimension.

In order to collect data from smartKYC servers, text data called snippets were collected separately for each frame of the entity documents to be collected using API calls with the help of entity review id. Snippets represent the basic information required for summarization. While collecting data, filtering was also applied on some features that came with the API response and were already

specified by smartKYC. For example, the source of the snippet text data and the reliability percentage of the source being more than 55 percent, not including some previously determined sources in the content data, not including text data tagged as 'not related' or 'irrelavent' on the server side, and the reliability rate of the date on which the data to be used to create the timeline was published being more than 55 percent. The purpose of determining the filters was determined to provide the balance between the desire to use comprehensive data and the desire to ensure information reliability.

Since tests were conducted with different models, the context length was initially kept as 9000 tokens, but it was made flexible in subsequent tests and left according to the tester's request. However, in order to keep the changes in the tests under control, the token length default value in the prototype remained as 9000. Although the context window size increases, if the entity review content size is huge, it will not be able to meet this need again and the content can be taken as much as the maximum content window size. Considering this constraint, a sequential approach was determined to add information from various sources. This method involved initially selecting the first snippet from each source, then adding subsequent snippets in order (for example, the second snippet from each source, then the third snippet) and continuing this iterative process until the cumulative token count approached the threshold. The main reason for this approach was to obtain a more objective approach by including information about the entity from various sources in the content and at the same time to prevent possible duplicate data from different sources. The collected snippets were then assigned to a JSON object to facilitate the creation of the content structure and possible subsequent filtering. Each snippet contained metadata such as the date the information was published, the source reliability percentage, the language in which the text was written, the related entity and the document provider. This structuring of the data facilitated subsequent progress and helped to form the content data that the model fed.

When a news source is thought to be talking about a specific event, person and those affected/influenced, it is seen that entity names are not always used in the content, instead pronouns such as he/she/they are used instead of names. However, since the reader has the ability to distinguish them and establish the connection between them, there may not be any confusion. On the other hand, if we consider this situation for an LLM, and also take into account the length of the content data, the desired result may not always be obtained. In order to eliminate possible ambiguities and improve the language model's understanding of the main subject, people and context, it was considered to

enhance pronouns with entity names. This involved replacing pronouns and partial names referring to the tracked entity with the entity's full name in accordance with the grammar rules of each particle's language. This step was crucial to ensure that the model correctly identifies and focuses on the primary subject, reducing misunderstandings caused by ambiguous references. With the help of linguists, after some tests to improve the summarization process, the content data was sorted by the publication date of the data to provide a logical flow of information. Although the initial publication dates were specified as the date of the event and caused hallucinations, the chronological order helped the large language model to better understand temporal relationships.The snippets that passed through this progression were brought together and became the final content data for the LLM.

3.3 Model Selection

Choosing an LLM that fits your task is a challenge for most projects, and is critical to creating an effective hashing structure for smartKYC. This process involves testing various LLMs, incorporating feedback from linguists, and ultimately selecting the model that best meets the project's goals in terms of accuracy, efficiency, and privacy.

3.3.1 Initial Experiments with Different LLMs

As mentioned in Chapter 2, most of the LLMs are pretrained models that can perform multiple tasks, including summarization. Depending on the task to be used, various methods such as fine-tuning, transfer learning and instruction learning can be integrated in the direction to be applied. While the length of this process and the ability to obtain accurate results are shaped according to the results obtained, the expected result may not be a hundred percent. In this project, the main goal was to obtain the desired structured output in line with the given content and instructions. For this, first of all, some LLMs were examined based on previous studies, considering the software and hardware requirements, summarization capabilities, compatibility with multiple languages and possible costs to the company. The evaluated models included GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, BART, Gemini, LLaMA2, Mistral, T5, Pegasus. Considering the mentioned interests, it was preferred to test Open AI models based on benchmarks. As mentioned before, among the tested models, Open AI models family has the best results, although they sometimes showed hallucinations.

However, after the improvement of the versioning of each Open AI LLM , each version of the model started to give better results than the previous one. For example, gpt-4 is better than gpt-3.5-turbo in both output structure and also the more correct results.

As mentioned before, among the tested models, Open AI models had the best results, although hallucinations were seen from time to time. However, after the updates of the LLM model, each model started to give better results than the other. For example, gpt-4 is better than gpt-3.5-turbo in both output structure and also the more correct results. At the same time, prompts continued to be improved for each frame in every problem seen with the help of linguist. Since it was seen in the literature that BART, Llama2 and Misral pretrained models require extra fine-tuning for summarization and since there is no such hardware and data set at the moment, in the first stage, only Llama2 was tested locally via huggingface transformers models. Although better results were obtained as the prompts were developed, in parallel, the latest updated Gemini 1 and Gemini 1.5 pro from Google models were also tested via API. However, when the results we obtained were examined, it was seen that there were problems especially in the output format and that it was insufficient to produce json structure.

First of all, although licensed models are intended to be used, there is always a concern about protecting users' information. Another applicable issue for this is anonymization, however, in the course of the project, open source models, Llama3 8B and Mistral 7B, have been tested on the AWS platform. The factors that LLMs should create consistent and concise summaries in the summary task, provide correct references to original snippets, obtain the expected output structure, and provide summaries without hallucination are important for model selection.

Based on the experienced models, it was seen that gpt-3.5-turbo and gpt-3 models produced fluent summaries but could not extract the desired details, while the gpt-4-turbo model provided results in the desired structure, which played a major role in this model taking its place in the testing stage. However, as seen in the tests conducted, it was also seen that it was better for the gpt-4o model to repeat the same result according to the same prompt. Tests on this are still ongoing and whether the gpt-4-turbo or gpt-4o model will take its place in production will depend on the test results. At the same time, cost calculations should also be taken into consideration when making a decision.

3.3.2 Human Reviewer Feedback

In order to be able to apply the metrics that are intended to be used to increase quantitative evaluations, such as BLEU and METEOR, reference summaries are needed. In the reviewed studies, the performance evaluations were generally provided using a general reference dataset. For example, there are datasets such as SummEval or WikiSum for the summary task. In these datasets, the training dataset consists of paragraphs, while the labels are human written summaries. However, since this project wanted to make the information collected on certain themes to be summarized in a structured format and the IDs of the reference snippets to be included in the output, it was deemed appropriate to prepare a test dataset for this project.

The steps in this process were as follows:

- Preparation of Reference Summaries: After data collection and compilation for a given entity, 50 reference summaries were prepared based on the LLM prompt and in the desired output format.
- Linguist review of prepared summaries: Prepared summaries were reviewed by a linguist who was knowledgeable about the subject matter to ensure there were no omissions and corrections were made.
- Evaluation Process: Summaries generated by Gemini,gpt-4-turbo-preview, gpt-4-turbo, and gpt-4o models were first compared with linguist aids, and then BLEU, BertScore, METEOR, and ROUGE scores were examined.

3.3.3 Selection of Best Performing Model

The evaluation, which combined both quantitative measurements and qualitative human feedback, found that while gpt-4-turbo is currently the model used in smartKYC, analysis showed that gpt-4o was noted for having better repetition and more consistent answers. No model changes were made in the production phase as testing is still ongoing.

The selection process began with each model producing summaries for a standardized set of documents representing the diverse and complex nature of smartKYC's data. These documents covered a variety of topics, including legal issues, financial records, and compliance reports, each with detailed terminology and complex structures. The models were evaluated on their ability to capture critical information, preserve the integrity of the original context, and include accurate references to the source material. Quantitative metrics

played a significant role in the evaluation, while assistance from linguists and process improvement contributed greatly. As mentioned earlier, while gpt-4-turbo is currently the model in active testing and demos, gpt-4o consistently outperformed other models across a variety of evaluation criteria:

- BLEU Scores: The gpt-4o and gpt-4-turbo models had very similar BLEU scores, while the gpt-4-turbo-preview model had a score almost twice that of the gpt-4-turbo-preview model. The gpt-4-turbo-preview model achieved higher BLEU scores, indicating greater overlap with human-written reference summaries in n-gram sequences. This shows that the model effectively captures key phrases and terminology used by human experts.
- ROUGE Metrics: With superior ROUGE-1, ROUGE-2, ROUGE-3 and ROUGE-L scores, gpt-4-turbo-preview demonstrated its ability to preserve the essential content and structural patterns of original documents. These metrics reflect the model's ability to summarize information without losing critical details.
- METEOR Scores: gpt-4o has the highest METEOR score which means that it is capable of precision and recall by taking into account synonymy and stemming, which are crucial for understanding and communicating the nuances of complex legal and compliance language.
- BERTScore: The superior BERTScore of the gpt-4-turbo-preview model, which is higher than 90 percent, highlights its ability to maintain semantic similarity to reference summaries, ensuring that the generated summaries preserve the intended meaning and context of the source material.

In addition to these metrics, evaluations were conducted to measure the proportion of correctly referenced snippets. Overall, models above gpt-4-turbo-preview were found to be successful in adding accurate references to the original snippets, which is important for users to backtrack the summarized information for verification, but this is still a work in progress. In particular, gpt-4-turbo and gpt-4 summaries were also found to exhibit a level of fluency and consistency comparable to human-written summaries. It was also discussed that gpt-4o and later models, which accept custom structured output formats for the output format, which is a common problem for all models, could also be used. However, it was requested that the Llama3 and Mistral models, which are still in the testing phase, be compared with gpt-x models.

Despite initial concerns about data privacy when using external APIs, strategies such as data anonymization are also on the agenda, as gpt-4-turbo, accessible via API, provides an instant solution with minimal integration challenges. Sensitive information in documents is also considered to be masked or replaced

with placeholders before being processed by the model, reducing the risk of confidential data exposure and ensuring compliance with privacy regulations.

In summary, the selection of gpt-4-turbo for the current project demo was based on an evaluation that took into account performance metrics, human feedback, and practical deployment factors. The model's ability to consistently produce high-quality summaries that are faithful to the source material, combined with its ease of integration and compliance with data protection requirements, made it the appropriate choice for the current smartKYC summary capabilities. The model is expected to significantly improve the efficiency and accuracy of information processing within the platform, and user testing will begin soon.

3.4 Prompt Engineering

3.4.1 Developing Prompts for Different Topics

The performance of LLMs in producing accurate and contextually relevant summaries is highly dependent on the design of the prompts used. Prompt engineering has therefore become a critical component in using LLMs for specialized tasks such as summarizing complex legal and compliance documents within smartKYC. This section outlines the development of specialized prompts for different topics, referred to as "frames," and discusses related methodologies and considerations.

Prompt engineering involves creating input instructions that enable LLMs to effectively produce desired outputs. The structure and content of a prompt can significantly impact the model's performance, especially in tasks that require precision and adherence to specific formats. In the context of summarization, prompts should be designed to yield concise, accurate, and consistent summaries that capture essential information while avoiding irrelevant details or hallucinations.

The project involved extensive study of prompt engineering techniques and linguistic assistance in identifying best practices, along with a review of relevant literature. Key resources include zero-shot, few-shot, and chain of thoughts prompts, which were observed to reduce hallucinations in the generated text. Understanding these techniques was important for designing prompts that could address the complexity and diversity of frames within smartKYC. Given the broad scope of topics covered by the frames (such as legal issues, bankruptcy cases, and lifestyle factors), it was necessary to develop customized prompts for each frame. For each topic, a detailed list was prepared outlining the following:

- Information Requirements: Key data points, specific topics, and details that the abstract should include.
- Restrictions: Issues that must be considered when creating the abstract.
- Format Specifications: Desired structure, language considerations, and inclusion of references.

This systematic approach ensured that the prompts were aligned with the specific objectives of each frame, leading to more targeted and effective summaries.

The experienced prompting techniques mentioned above are also listed below with their definitions.

- Zero-Shot Prompting Technique: Providing the model with only example-free task instructions. This method is simple but may not yield the best results for complex tasks.
- Few-Shot Prompting Technique: Including examples of desired inputs and outputs in the guidance to guide the model. While this can improve performance, it has been found to increase the risk of hallucinations due to the long and complex nature of the input documents.
- Chain of Thought Prompting Technique: It encourages the model to produce intermediate reasoning steps before producing the final summary. This method has been shown to improve the model's ability to handle complex tasks and produce more accurate results.

At the same time, studies have shown that experiments with a few-shot prompting technique that providing examples in the content leads to hallucinations, and the model generates information that is not present in the source material. This problem is exacerbated by the length and complexity of the input texts, as the model's attention is divided between processing the examples and the actual content to be summarized. In contrast, the chain of thought prompt yielded better results. Directing the model to express intermediate reasoning steps helped it process information more effectively and produce consistent and accurate summaries. This approach allowed for improved inclusion of relevant details.

The prompts were structured in three distinct stages:

- Preamble: The main goal and context were explained to the model. This section clarified the nature of the snippets provided and any special formatting, such as identifiers or publication dates.

- Body: Detailed instructions were provided on how to create the summary. This included specifying the information to be included, the required categorizations, and how to handle various data points. Constraints were also outlined to prevent irrelevant information from being included.
- Closing: The desired output structure was specified, including formatting requirements and any additional instructions. This section ensured that the model's output would be in a usable format, facilitating integration into smartKYC's system.

Specific examples of the prompts are provided in Appendix A.

3.4.2 Optimization of Prompts Based on Results

Developing effective prompts is an iterative process that involves continuous testing and improvement. This subsection details the methods used to optimize prompts based on the results obtained from the model outputs. For each prompt, a series of sequential tests were conducted to evaluate the model's performance. The outputs were analyzed for:

- Accuracy: Whether the summaries included all critical information according to the requirements.
- Consistency: Flow of the summaries and readability of summaries.
- Constraint Compliance: Adherence to specified guidelines and exclusion of prohibited content.
- Format: Correctness of output structure, including appropriate use of identifiers and references.

However, common issues identified by the linguist during testing include:

- Omission of Key Details: Important information was sometimes missing from summaries.
- Inclusion of Irrelevant Information: The model occasionally added irrelevant content or failed to exclude prohibited details.
- Formatting Errors: Deviations from the specified output structure occurred and affected the usability of the summaries.

To address these issues, the prompts were revised to provide clearer instructions and strengthen restrictions. This included restating the instructions, emphasizing critical requirements, and adjusting the level of detail in the prompts. Encouraging the model to use chain of thought reasoning was helpful. The

results were improved by asking for more comprehensive information by instructing the model to consider the steps required to create the summary.

To reduce hallucinations, the prompts were adjusted to focus the model's attention only on the provided snippets. Instructions were added to prevent the inclusion of information not included in the snippets and to verify the facts before including them in the summary. These adjustments reduced the generation of fabricated content. It is also important to strike a balance between the length and complexity of the prompts. Overly long or complex prompts can confuse the model and reduce its effectiveness. Therefore, prompts should be prepared to be as concise as possible while providing all necessary instructions, and negative prompts should be avoided as much as possible.

Through iterative refinement, the prompts were optimized to produce summaries that met the project's goals and are still being improved with the help of the linguist. The final prompts resulted in:

- Improved content in the summaries: Summaries contained most critical information and adhered to constraints.
- Improved Consistency: Summaries had better logical flow and readability.
- Consistent Formatting: Outputs followed the specified structure, facilitating seamless integration into the smartKYC system.
- Reduced Hallucinations: There were fewer instances of fabricated or irrelevant information.

Prompt engineering was a critical factor in the successful implementation of the summarization system. By developing specific prompts for each framework and systematically optimizing them based on test results, significant improvements in project summary quality continue to be achieved. Examples of final prompts are provided in Appendix A.

3.5 Summarization Process

The summarization process is a component of the project that integrates the data collection, model selection, and prompt engineering practices described in previous sections. This process involves creating summaries from the collected data using the selected language model and then refining these summaries to produce concise and comprehensive overviews. This section covers the methodologies used to create the initial summaries, followed by general summaries based on the dimensions of these summaries.

3.5.1 Generating Summaries

The generation of summaries is a multi-step process that leverages the prepared data, optimized prompts, and the selected language model to produce wanted summaries for each frame within the smartKYC platform.As detailed in Section 3.2, the input data for summarization in this project consists of collecting and preparing snippets based on specific frames associated with a review. These snippets are structured in a JSON format that contains not only the text but also metadata such as publication dates, source credibility scores, languages, and related entities, but this JSON structure is used only to organize the content that will be included in the prompt. Before feeding this data to the model, additional preprocessing steps are performed as follows:

- Entity Normalization: To ensure clarity and reduce confusion in the generated summary, ambiguous pronouns, pronouns used instead of nouns, are replaced with the full name of the entity being tracked.
- Chronological Order: Since the chronological ordering of the plots in the tests improved the resulting summary, ordering the snippets by event dates helps the model understand temporal relationships to maintain a logical flow of information.
- Excluding Information: Snippets that comes from sources which are not reliable are excluded from the content.

Using the prompts developed and improved in Section 3.4, the model is instructed to generate summaries that meet certain criteria. Among the prompts, the focus is on important events, facts, and relevant details, and what information should be included, while clearly explaining what information should not be included. At the same time, the model is encouraged to process the information in a logical order by explaining the desired output structure and the references that are desired to be included, and the consistency of the summaries is increased.

LLM is used to create summaries based on prepared input data and optimized prompts. The process includes:

- Model Call: The Model API is called with the input data and prompt.
- Create Summary: The model produces a summary that is expected to be consistent, accurate, and formatted according to the specified guidelines.
- Include References: Increases traceability by adding references to the original snippets using the model identifiers, according to the instructions in the prompt.

Once the summary is created, it goes through a validation process to ensure it meets the required standards:

- Format Validation: Automated checks verify that the summary follows the specified JSON structure and formatting rules.
- Content Validation: The abstract is reviewed to ensure that it contains all critical information and that references accurately correspond to the original snippets.
- Language and Clarity Assessment: The text is reviewed for grammatical correctness, consistency, and readability.
- Any issues identified during validation are addressed by improving the prompt or preprocessing steps and recreating the summary as necessary.

Several difficulties were addressed during the summarization process. The first of these was that not all available content could be retrieved as content due to the context window of the model, so the input snippets had to be carefully selected and arranged, as explained in Section 3.2. Another was that the model could provide information that was not in the content and that this information would be referenced by the snippets in the content. Finally, consistency was required for each generated summary.

3.6 Human Written Summaries

Human-written summaries play an important role in both the development and evaluation of automated summarization systems. Although open source datasets are available for system evaluation, the points required for sufficient accuracy by each system are different. Therefore, it is important that human-written summaries are created in accordance with the rules that the model must follow and in line with the same output format. This approach serves as an important standard against which the performance of LLMs can be measured and provides a measure of accuracy, consistency, and relevance. This section describes in detail the process of creating these summaries and provides information about their key role in evaluating the effectiveness of language model outputs in the smartKYC platform.

3.6.1 Creation Process

The generation of human-written summaries provides a reliable basis for evaluating the performance of the language model. Summaries were generated

for 50 different frames owned by a given entity. When summarizing, the instructions in the prompts sent to the language model were followed exactly, aiming to ensure consistency in the goals and constraints guiding both the human- and LLM-generated summaries.Without access to the language model output, the human summary writer only considered these snippets and instructions to create their summaries. This blind-writing process was important to avoid any bias and ensure an unbiased comparison between the human- and LLM-generated summaries.

The key events and basic information about the entity were organized according to the instructions in the prompt. The human writer prepared the summaries according to the same rules, as the instructions specified that references to the original snippets be included using their identifiers, that factual accuracy be maintained, and that the specified output structure and formatting requirements be adhered to. Once completed, human-written summaries were reviewed by the linguist and project manager. This quality assurance step is to ensure that all summaries meet the high standards set for the project and that any inconsistencies are resolved.

The rationale behind this process is to ensure that human-written summaries provide a consistent and reliable basis for evaluation. By using the same instructions and data, and avoiding any influence from the output of the large language model, summaries reflect independent interpretations of the data. This consistency is crucial to ensuring a fair and objective comparison between human- and machine-generated summaries.

3.6.2 Role in Evaluation

Human-written summaries are used to evaluate the performance of automated summarization systems. They can be thought of as reference summaries that embody the nuances of human understanding and language use that the models attempt to emulate. Human-written summaries provide a benchmark for both quantitative and qualitative evaluations of the language model's output.

In quantitative assessment, human summaries are used as ground truth references to calculate statistical assessment metrics such as BLEU, ROUGE, METEOR, and BERTScore. These metrics provide objective measures of model performance by measuring the similarity between model-generated summaries and human-written summaries. For example, the ROUGE score evaluates the overlap of n-grams, allowing for assessment of content coverage and the model's ability to capture key information found in human summaries. Similarly, BERTScore

uses contextual embeddings to measure semantic similarity, providing insights into the model's ability to preserve the meaning of the original text.

Qualitative assessment involves analyzing the readability, consistency, and factual accuracy of model-generated summaries by comparing them to human-written summaries. This process evaluates whether the model captures critical information, maintains logical flow, and follows specified guidelines. Inconsistencies across summaries highlight areas where the model underperforms, such as omissions of important details or inclusion of inaccuracies. This analysis facilitates targeted improvements in model performance. Human-written summaries can guide the development of model outputs by providing concrete examples of high-quality summaries that the model can aim to replicate. Demonstrating that a language model can produce summaries that closely match human-written summaries validates its effectiveness and relevance in practical applications. However, it is important to acknowledge the inherent challenges of using human-written summaries for evaluation. Summarizing involves subjective decisions about which information is most important, and different people may produce different summaries from the same set of pieces. This variability should be taken into account when interpreting evaluation criteria and comparing model outputs to human references. Ensuring the quality of human summaries is also crucial; the reliability of the assessment depends on the summary creator's adherence to the guidelines and knowledge of the subject matter.

An assessment of the LLM performance was conducted by integrating human-written summaries into the assessment framework. The summaries provide both quantitative measures and qualitative insights, allowing for direct comparison with model-generated outputs. Feedback from this assessment drives iterative improvements to the summarization process, including adjustments to prompts and preprocessing steps. The results are satisfactory, but judging by the assessment metric results seen thus far, there is room for improvement.

Ultimately, human-written summaries are an important part of the development and evaluation of automated summarization systems. They serve as a goal for models to achieve and a tool to measure progress. The creation and use of these summaries in this project provided a reference for evaluating and improving the language model's summarization capabilities.

4

Experiments and Results

4.1 Initial Model Comparisons

The first phase of the experimental study deals with comparing various large language models to determine the most suitable one for the summarization task on the smartKYC platform. The evaluated LLMs included gemini-1.5-pro, gemini-1.0-pro, gpt-4-turbo-preview, gpt-4-turbo, gpt-3.5-turbo LLMs. Each model was evaluated on its ability to produce accurate, consistent, and contextually relevant summaries that closely match human-written summaries. The selection process for the models was guided by their respective capabilities to handle complex natural language processing tasks, especially in the area of summarization.

Considering the examined LLMs, gemini-1.5-pro and gemini-1.0-pro developed by Google are pretrained models designed to manage complex language understanding and generation tasks. Its advanced architecture shows potential suitability for summarizing complex documents with domain-specific terminology. The Open AI family models are a series of language models optimized for faster inference while aiming to maintain high-quality outputs. Its design prioritizes efficiency, which can be useful for processing large amounts of data in real-time applications. The LLM model gpt-4o, which has recently been released to the user experience from the Open AI family models, is an optimized version of the gpt-4 model designed for improved consistency and repeatability in output. Considering the importance of reliable and repeatable results in compliance contexts, and as a result of project manager tests, gpt-4o is a promising LLM. However, tests are still ongoing with updated prompts.

The experimental evaluation involved providing each model with the same

set of input data and prompts. The input data consisted of snippets associated with specific frames that were pre-processed and structured as described in the methodology section. These snippets contained complex legal and financial information that required the models to understand and accurately summarize complex content.

To ensure an unbiased assessment, human summary writers were used to create reference summaries for the models. This approach allows for a fair comparison of the models' inherent ability to interpret and summarize the data provided based solely on the input and prompts, allowing for the assessment of the models' performance on the summarization task.Summaries written by human writer are an important part of the performance evaluation of models, helping to identify their strengths and weaknesses.

When Llama2 was tested on a local computer, it could be said that the summaries it generated were based on the given content structure, but an output in the desired format could not be obtained. The first stage was also found to be insufficient because its performance was slow due to high hardware requirements and because it made general summary inferences far from the level of detail.

While the Google models gemini-1.0-pro and gemini-1.5-pro produced summaries with reasonable accuracy, the output structure exhibited inconsistencies and hallucinations. While some summaries were satisfactory, others contained irrelevant information or were inconsistent in referencing the original snippets. This inconsistency raises concerns about the reliability of the model, and since inconsistent information can lead to important conclusions, the Google models also fall short of the intended structure.

Considering the Open AI family, the models generally show strong language generation capabilities by producing fluent and consistent summaries. However, some variability was observed when statistical methods were used. However, comparisons are still ongoing between gpt-4o, gpt-4-turbo and gpt-4-preview, as consistent and repeatable outputs are desired.

In addition, the Llama3 8B and Mistral 7B models, which are models that the Amazon Web Services (AWS) platform allows to be deployed, were added to the server where the project prototype is located with the help of AWS APIs, and work has begun to ensure their testing.The initial model comparisons highlighted Open AI gpt-4 series as the most suitable models for the summarization task within the smartKYC platform. Its superior performance in generating accurate, coherent, and consistent summaries, coupled with its ability to handle complex, domain-specific content, made it the optimal choice. For choosing which model of family is more suitable there are some more tests which are still

work on. However, although it is not based on any evidence, it has been seen in the tests that gpt-4o generates texts that are more suitable for the desired output.

4.2 Prompt Optimization Results

The performance of a language model in producing accurate and consistent summaries is greatly affected by the quality of the prompts used to prompt it. Therefore, after each test result, the prompt optimization process continues using the prototype and with the help of linguists. This process involves close collaboration with linguists and an iterative process of improving the model based on its performance analysis.

Prompt optimization begins with a detailed analysis of the summaries produced by gpt-4-turbo-preview LLM, which provides the most stable outputs. During the analysis, issues that needed to be addressed to improve the quality of the outputs included duplicate summaries, empty or incomplete summaries, inclusion of hallucinatory dates (taking the document publication date as the event date), missing details, and the presence of information missing from the snippets. These issues led to concerns that the prompts were vague or insufficiently detailed, which may have led to unexpected behavior from the model. In the process, the prompts were improved by addressing the listed issues.At the same time, the contribution of the linguist was significant, as he provided valuable insights into how the prompts could be adjusted to mitigate the identified problems. It is important to understand the nuances of language that may affect the model's interpretation of the prompts. It is necessary to analyze the wording, structure, and content of the prompts to eliminate potential confusion or misdirection.

The optimization process is iterative. Each iteration involves modifying the prompts to target specific problems and then testing the model outputs to evaluate the effectiveness of the changes. For example, when the model produced summaries with frequently repeated mandates and themes, the prompt was revised to include explicit instructions to avoid unnecessary repetition. Similarly, empty summary outputs led to the model emphasizing that all relevant information contained in the provided snippets should be included. This cycle of refinement and evaluation continues, with each adjustment informed by observed results. The iterative nature of the process allows for the prompts to be progressively refined, tailored to effectively extract the desired responses from the model. This approach also demonstrates that the changes are based on empirical evidence of what works rather than theoretical assumptions.

The prompt optimization process provided significant improvements in model performance. One major improvement was the elimination of unnecessary repetition in summaries. By clarifying the instructions and emphasizing the need for conciseness, the model began to produce more fluent summaries that conveyed essential information without unnecessary repetition.

The problem of empty or incomplete summaries was addressed by revising the prompts to explicitly instruct the model to include all relevant details. By ensuring that summaries were comprehensive and informative, it was able to extract information from the required details of the snippets. This adjustment reduced the occurrence of empty summaries.

A significant challenge was hallucinatory dates, where the model provided incorrect or fabricated dates. To alleviate this, the prompts were improved to provide clear guidance on how to process dates. The model was instructed to rely solely on dates found in snippets and to prioritize event dates over publication dates when summarizing information. This clarification reduced the inclusion of incorrect dates, increasing the factual accuracy of the summaries.

Another improvement is customizing prompts to specific frames or topics. By customizing the prompts to the context of each frame and adding details, the model is given clearer content and the ability to produce relevant summaries across different topic areas. This customization, specifying the types of information to focus on and the appropriate terminology to use, allows the model's output to be more closely aligned to domain-specific needs.

Additionally, the issue of the model including information in the snippets of the summaries was also investigated. In the prompts, it was emphasized that the summaries generated by the model were based on the snippets and constraints provided in the content data and that any additional information not included in the provided data was avoided, thus reducing hallucinations. At the same time, a clear structure was provided regarding the desired structure of the summaries and restrictions were imposed on the output format generated by the model. The main topics and subjects that should be focused on for each frame in the prompt were also clearly stated. Despite the improvements, some problems were still found. One of the problems was that it cited false positives and sometimes referenced snippets that were not relevant or correct. This affected the traceability and reliability of the summaries. Considering that it was necessary to verify the relevance of each referenced snippet, the model was informed that only snippets that directly supported the content of the summary should be included as references.Another problem encountered is inferring an individual's origin without sufficient evidence from the snippets, or hallucinations resulting

from the language used. Given that each model has different levels of knowledge in different languages, it is clear that most models generally give more efficient answers in English. To prevent such unfounded inferences, the model was instructed in the prompts to avoid making assumptions about nationality unless explicitly stated in the data, and was also given the insight that the language used could be a language other than English.

Another problem encountered was that the occasional summary in languages other than English was produced without sufficient evidence from the snippets, which was a challenge, especially since the instructions stated that the output should be entirely in English. The language requirement was developed in the prompts, and the instructions reminded the reader to translate any non-English text into English before processing, and to produce the summary only in English. This helps to ensure consistency in the language of the output.

The optimization process resulted in a refined prompt structure that effectively guided the model to produce high-quality summaries. The prompts were carefully crafted to balance clarity, specificity, and brevity, ensuring that the model received all necessary instructions without being overwhelmed by complexity. The prompts were structured into three main components:

- Preamble : This section sets the context for the task and provides the model with an understanding of the input data and special formatting considerations. For example, it explains that snippets are delimited by triple back-ticks and that each snippet is preceded by an identifier. It also explains how publication dates are presented and instructs the model on how to handle nonstandard elements.

- Body : The body provides detailed instructions on how to create the summary. It specifies information to include, such as key events or facts about the entity being tracked, and outlines any restrictions or guidelines that must be followed. These include instructions for handling dates, avoiding repetition, and ensuring that content is based on the provided snippets. The body also addresses formatting requirements, such as the structure of the output and the inclusion of references.

- Closing : The closure reinforces expectations for the response, summarizes key points, and defines the structure of the response. It reminds the model of the language requirements and other critical issues by specifying the desired output structure, usually in JSON format. The closure ensures that all aspects of the prompt are consistent and that the model clearly understands what the expectations are.

By optimizing each component of the prompts, the model's performance continues to improve. The prompts are intended to be more effective in producing accurate, consistent, and relevant summaries that are compliant with

the requirements of the smartKYC platform. Collaboration with linguists and the iterative nature of the optimization process are instrumental in achieving these improvements. Examples of improved prompts that demonstrate specific adjustments and the positive impact they have on the model's output are provided in Appendix A. These examples are intended to demonstrate how precise and well-structured prompts can significantly impact the quality of the summaries produced by the language model, emphasizing the importance of prompt engineering in natural language processing applications.

4.3 Final Summarization Results

While the prompt optimization and summarization methodology is still being improved, the final summaries generated by the gpt-4o, gpt-4-turbo and gpt-4-turbo-preview models were subjected to a comprehensive assessment to assess their quality and effectiveness within the smartKYC platform. This assessment aims to determine the extent to which improvements in prompt optimization and model fine-tuning are reflected in improved performance in generating summaries that meet the requirements of the platform.

The assessment of summaries is based on several key criteria: accuracy, consistency, completeness and formatting compliance. These criteria are considered to reflect the fundamental qualities required for effective summarization in the context of compliance and due diligence reporting.

Accuracy is an assessment of how faithfully the summaries represent the information contained in the original snippets. This includes checking for factual accuracy, ensuring that all statements in the summaries are supported by the source material, and verifying that there is no misinformation or hallucination.

Consistency focuses on the logical flow and readability of the summaries. A coherent summary presents information in a clear and logical order that makes it easy for the user to understand. This criterion also takes into account the use of appropriate language, sentence structure, and avoidance of ambiguity or confusion.

Completeness evaluates whether the summaries include all critical information about specific frames. Summaries are reviewed to ensure that they cover all key details provided in the snippets, without skipping over important points that could impact the user's understanding or decision-making process.

Formatting Compatibility is critical for seamless integration of the results in the system. Summaries must strictly adhere to the specified output structure, including correct use of JSON formatting, appropriate inclusion of identifiers,

and correct referencing of the original snippets.

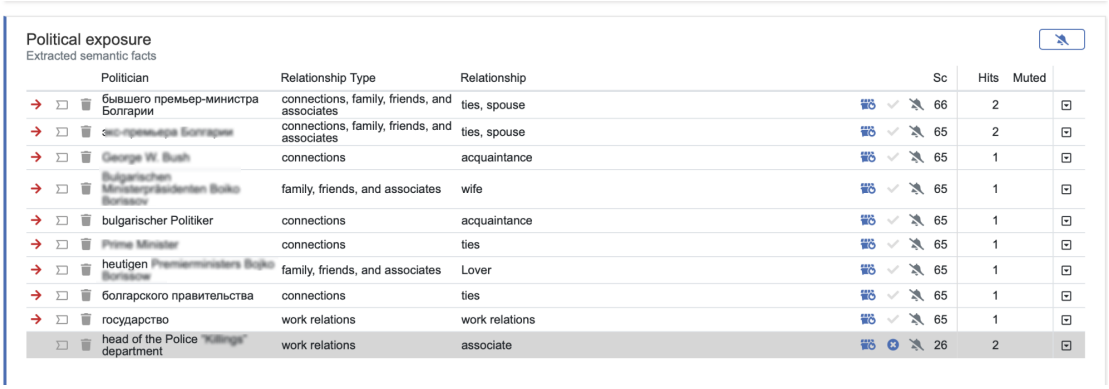

Figure 4.1: smmartKYC platform Political Exposure Initial format

The final summaries show significant improvements in all evaluation criteria, indicating that rapid optimization and methodological improvements have a positive impact on the model's performance. In terms of accuracy, the summaries are able to accurately reflect the content of the snippets, and a decrease in hallucinations has been observed. The model is able to effectively capture essential information without introducing unsupported or erroneous statements. In terms of consistency, the language used in the summaries is clear and logically sound, making the summaries more readable and user-friendly. The information is presented in a structured manner, making it easier to understand. Improvements in sentence structure and elimination of unnecessary repetition contributed to the clarity and conciseness of the summaries. For completeness, the summaries generated by the model consistently contain important details about each frame. The summaries comprehensively cover critical information from the snippets, minimizing omissions. Work continues to ensure formatting compatibility so that the outputs adhere to the required JSON structure with the correct use of identifiers and references. For this purpose, the response format method supported by gpt-4o and later models is considered. However, it is still in the testing phase. Sometimes, the model's response format is not as expected, but this can be corrected by processing the response.

The final summarization results highlight the success of prompt optimization and methodological improvements in improving the performance of the LLMs. The model demonstrated significant improvement in producing summaries that are accurate, consistent, complete, and compliant with formatting requirements. These improvements significantly contribute to the usability of the smartKYC platform, providing users with high-quality summaries that support effective compliance and due diligence processes. The combination of qualitative assessments and quantitative metrics provides a comprehensive val-

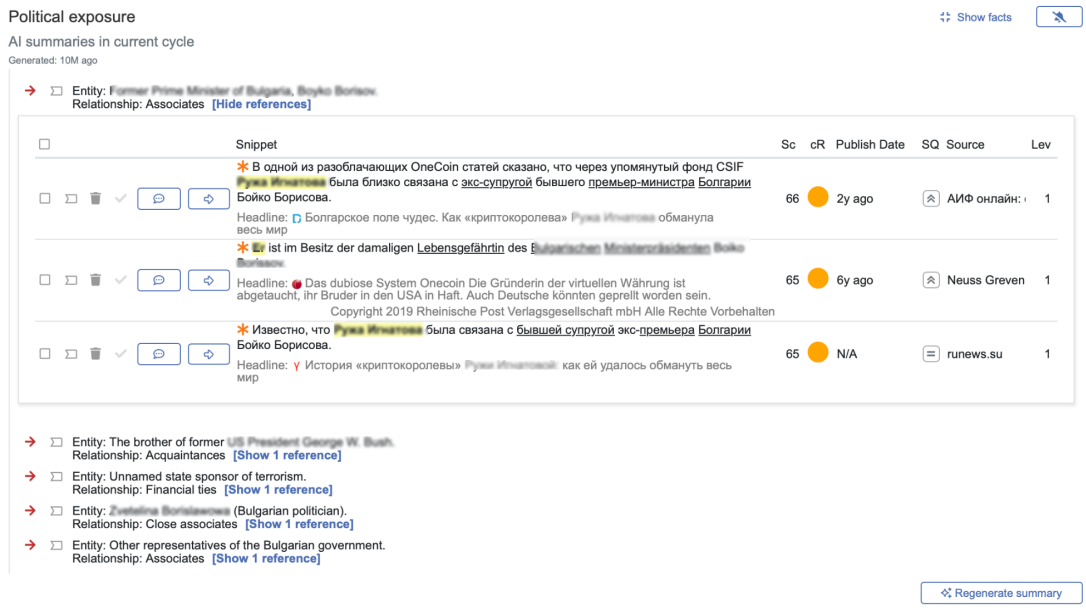

Figure 4.2: smmartKYC platform Political Exposure Final format

idation of the model's capabilities. The improved performance highlights the importance of prompt engineering and iterative improvement when leveraging large language models for specialized tasks. The observed positive results provide a strong foundation for continued implementation and development of the model within the smartKYC platform, with potential for further improvements through ongoing research and refinement.

4.4 Evaluation with Human-written Summaries

In order to evaluate the performance of LLM, a comparison needs to be made between the summaries generated by the model and the human-written summaries detailed in Section 3.6. This evaluation aims to assess the extent to which the models can replicate the quality, accuracy, and depth of understanding demonstrated by human summary writers in summarizing complex legal and compliance documents. However, using these evaluation metrics, it is aimed to identify both the strengths and weaknesses of the model, which is essential for the summarization system.

The evaluation process involved a systematic approach to ensure fairness and objectivity. Each model-generated summary was matched to its corresponding human-written summary by aligning it to the same input snippets and frames. This one-to-one matching helps to make the comparison precise by minimizing variables that could affect the evaluation.

The quantitative analysis was conducted using established evaluation met-

rics including BLEU [63] , METEOR [64] , ROUGE [65] and BERTScore [66] . These metrics helped to quantify the similarity between the model-generated summaries and human-written references. BLEU score focuses more on accuracy and evaluates the overlap of n-grams. METEOR examines the similarity between texts by considering synonyms and morphological variations by considering whole-word matches. ROUGE examines the similarity ratios by evaluating the overlap of n-grams and longest common sub-strings. BERTScore evaluates the similarity ratio between summaries using contextual embeddings to measure semantic similarity.

In addition to quantitative metrics, qualitative analysis was conducted with the assistance of a linguist for legal and compliance issues. This assessment evaluates the summaries for accuracy, consistency, completeness and compliance with the specified guidelines and formatting requirements. The model was evaluated to see if it effectively captured key information, maintained a logical flow, and presented content in a manner consistent with professional standards.

When comparing the Open AI family models, the comparison between the summaries generated by the gpt-4-turbo-preview model and the human-written summaries showed a high level of agreement in terms of content, structure and quality, and achieved better results than the gpt-4-turbo and gpt-4o models. However, it needs to be tested again after the prompt improvements made in the next process. Because the results of the gpt-4-turbo-preview and gpt-4o models are close to each other as can be seen in the summaries. Quantitatively, the model's summaries showed superior performance in all evaluation criteria. The model achieved a BERTScore score of 0.9335904762, indicating a high level of semantic similarity with human-written summaries. The BLEU score of 0.3626554529 indicates that the model's word choices closely match those of human writers and reflects its sensitivity in n-gram usage.

The METEOR score of 0.7341786968 hit the effective recall and agreement with human expressions, including synonyms and morphological variations, while the gpt-4o model performed better with a METEOR score of 0.7401935994. The ROUGE metrics further highlighted the model's performance with ROUGE-1 being 0.7346381245, ROUGE-2 being 0.5572393965, ROUGE-3 being 0.4231505812, and ROUGE-L being 0.7165446162. These scores indicated an overlap between the summaries generated by the model and human-written references in unigrams, bigrams, trigrams, and longest common subsequences.

Qualitatively, the summaries of the gpt-4 model family were generally comparable to human-written summaries in terms of fluency, readability, and adherence to guidelines. The model was able to capture key information by accurately

reflecting the content of the provided snippets. The logical flow and structure of the summaries were consistent, and the inclusion of appropriate references to the original snippets increased followability. While the model accurately summarized explicit information from the snippets, it occasionally lacked the nuanced interpretation and deeper contextual understanding that human annotators bring to their summaries.

For example, in complex legal cases, human summary writers may notice the broader implications of a particular event or identify underlying themes not explicitly stated in the snippets. Working solely from the provided data and instructions, LLM tends to be more realistic in its summarization, focusing on explicit information without inferring additional context.

Evaluation with human-written summaries highlighted both the strengths and weaknesses of LLMs. The model was observed to be accurate, consistent, and prone to produce summaries that are consistent with human writing in terms of linguistic expression and adherence to instructions. However, the observed differences in nuanced interpretation and contextual understanding highlight the limitations of existing language models. The LLM's reliance on explicit knowledge without the ability to incorporate external knowledge or make unspecified inferences means that it cannot fully replicate the depth of insight provided by a human summary writer. This limitation is inherent in LLMs, which, despite their advanced natural language processing capabilities, lack the experiential learning and abstract reasoning capacities that humans possess.

The resulting evaluation shows that the model is a capable model for automated summarization that is close to human performance in many aspects. Its strengths in accuracy and consistency make it a suitable choice for deployment on the smartKYC platform. However, as mentioned earlier, gpt-4o model is considered to be more efficient in terms of repeatability of the given output, and therefore, testing is ongoing.

Future work could focus on improving the model's ability to incorporate contextual understanding and domain knowledge, potentially through techniques such as incorporating retrieval-based information or training on more comprehensive domain-specific datasets. Additionally, developing mechanisms for human-in-the-loop interactions where human evaluators can easily augment or improve the model's output could further optimize the summarization process.

4.5 Analysis of BLEU, METEOR, ROUGE, BERTScore Metrics

To quantitatively evaluate the performance of gpt-4-turbo-preview, gpt-4-turbo and gpt-4o, standard evaluation metrics were employed. These metrics provide a numerical basis for comparing the model-generated summaries with human-written references.

- BERTScore: It utilizes contextual embeddings to evaluate semantic similarity between the generated summary and the reference with providing a more nuanced assessment.
- BLEU Score: It measures the overlap of n-grams between the generated summary and the reference summary with focusing on precision.
- METEOR Score: It considers exact word matches and accounts for synonyms and morphological variants with emphasizing recall.
- ROUGE-1: It measures overlap of unigrams with emphasizing recall.
- ROUGE-2: It measures overlap of bigrams with emphasizing recall.
- ROUGE-3: It measures overlap of trigrams with emphasizing recall.
- ROUGE-L: It measures overlap of the longest common subsequence with emphasizing recall.

Table 4.1: Evaluation Metrics for GPT-4 Turbo and GPT-4o

Metric	**gpt-4-turbo-preview**	**gpt-4-turbo**	**gpt-4o**
BERTScore	0.9335904762	0.865037931	0.8775928571
BLEU	0.3626554529	0.1559632544	0.1700634505
METEOR	0.7341786968	0.565444583	0.7401935994
ROUGE-1	0.7346381245	0.5318764115	0.7085358643
ROUGE-2	0.5572393965	0.3442254247	0.5063149974
ROUGE-3	0.4231505812	0.2203992926	0.3537146449
ROUGE-L	0.7165446162	0.4867269249	0.6841406734

- BERTScore: gpt-4-turbo-preview achieved a higher BERTScore, indicating better semantic similarity with the human-written summaries. The higher score reflects the model's superior ability to capture the meaning of the content.
- BLEU Score: gpt-4-turbo-preview's BLEU score was higher, demonstrating greater precision in matching the reference summaries' n-grams. This suggests that this model generated summaries with word choices more closely aligned with human writers.
- METEOR Score: Although significant improvement in the METEOR score for gpt-4o indicates better recall and alignment with human expressions, including synonyms and morphological variations, gpt-4-turbo-preview model is also as well as gpt-4o.
- ROUGE-1 and ROUGE-2: gpt-4-turbo-preview showed substantial gains, reflecting better overlap in unigrams and bigrams with the reference summaries.
- ROUGE-3: The higher score for gpt-4-turbo-preview suggests improved capture of longer phrase structures.
- ROUGE-L: The increased score indicates that gpt-4-turbo-preview's summaries share longer common subsequences with the human references, demonstrating better structural alignment.

Although the test are still continuing, the superior performance of gpt-4-turbo-preview across all metrics confirms for this test its effectiveness in generating high-quality summaries that closely resemble human-written ones. The higher scores indicate that the model not only captures the essential content but also replicates the linguistic patterns and structures used by human annotators.

- Proportion of Referenced Snippets (Higher = Better): 0.88
- Average Over-referenced Snippets per Frame (Lower = Better): 0.53

Table 4.1 showed that for applied test gpt-4-turbo-preview effectively included relevant snippets in the summaries while minimizing the inclusion of irrelevant references. A high proportion of correctly referenced snippets enhances the utility and reliability of the summaries for users who require traceability to the source material.

In a test with 50 human-written summaries, quantitative analysis using BLEU, METEOR, ROUGE and BERTScore metrics shows that gpt-4-turbo-preview performs better in producing summaries closer to human-written references. Improvements in these metrics reflect improvements in accuracy, consistency and semantic similarity. Combined with the positive results from qualitative

assessments and on-the-fly optimization, this is the model that is suitable for the system in the tests conducted so far. However, since the prompts and the system continue to be developed, the selected model is open to change based on new test results.

5

Further Improvements and Future Work

5.1 Fine-tuning a Custom Language Model

5.1.1 Approach and Methodology

Building on the foundational work described in the previous sections, a key avenue for further improvement involves fine-tuning the studied language models, particularly the gpt-4-turbo-preview and gpt-4o models, to improve the summarization capabilities in the smartKYC platform. This approach aims to take into account the limitations associated with relying on external APIs, such as data privacy concerns and dependency on third-party services, while also achieving repeatable results.

To this end, we have started working with open-source language models, particularly LLaMA 3 8B and Mistral 7B, due to their effectiveness in natural language processing tasks and their suitability for fine-tuning. To facilitate scalable and efficient training, both models are supported by Amazon Web Services (AWS) cloud platforms. AWS provides robust computational resources, including GPU-accelerated instances, which are required to train large-scale language models. This cloud-based approach allows for flexible resource allocation and management required to meet the computational demands of fine-tuning.

An integral component of the fine-tuning involves assembling a comprehensive dataset representing the domain-specific language and content encountered within smartKYC. Over the past year, we have systematically collected data that includes snippets and summaries created during standard transactions, human-

written summaries discussed in Section 3.6, and documents annotated with metadata such as entity names, publication dates, source reliability scores, and thematic classifications.

The data collection process includes cleaning and normalization to remove noise and correct formatting issues, ensuring consistency across the dataset. To maintain data privacy standards, personally identifiable information can be anonymized using techniques such as pseudonymization and entity masking to prevent sensitive data from being disclosed. The data is then encoded in a format that complies with the input requirements of the LLaMA 3 and Mistral 7B models. The data is structured into input-output pairs suitable for supervised fine-tuning, where inputs are chunks and outputs are corresponding summaries.

The fine-tuning process starts with pre-trained weights of the selected language models using the available language knowledge. Training configurations such as learning rate, batch size, number of epochs, and optimization algorithms should be determined based on preliminary experiments and best practices. Appropriate loss functions such as cross-entropy loss guide the training process to produce accurate summaries. The training cycle feeds input-output pairs to the model, allowing it to learn the mapping from snippets to summaries. Validation sets monitor performance and prevent overfitting by using early stopping mechanisms when necessary. Model performance can be evaluated using metrics such as BLEU, ROUGE, METEOR, and BERTScore to ensure agreement with human-written summaries.

Among the technical aspects, cloud platforms such as AWS GPU can be used to provide the necessary computational power for training. Techniques such as mixed-precision training and gradient checkpointing optimize memory usage. If necessary, distributed training across multiple GPUs or instances can speed up the fine-tuning process. All training procedures must comply with ethical guidelines and data protection regulations such as GDPR.

5.1.2 Expected Outcomes

Fine-tuning custom language models is expected to provide several key benefits. By tailoring models to our specific domain and data, improved summarization performance is expected. Fine-tuned models can produce summaries that are more closely aligned with human-written summaries in terms of content and style. Improved consistency in output across topics and frameworks and reduced variability in summary quality are also goals. Processing data in our controlled AWS environment will address privacy concerns by eliminating

the need to transmit sensitive information to external services. Data protection compliance is facilitated by maintaining tight control over data processing processes. Eliminating reliance on third-party APIs can lead to significant cost savings over time. Cloud platforms like AWS enable scalable resource management and optimize operational costs based on demand.

However, it is important to acknowledge the resource-intensive challenges of fine-tuning large models. Efficient resource utilization and the potential use of smaller, optimized models can alleviate this problem. This trial-and-error process requires ongoing maintenance and updates to ensure models are performing optimally, and may require the establishment of a team for model management to meet these operational demands.

5.2 Potential Research Directions

Attempting to fine-tune custom language models opens up a number of avenues for future research and development. Exploring advanced models by investigating the potential benefits of using larger models, such as LLaMA 2 with 8B or 70B parameters, can improve performance, offset by increased computational costs. Using model compression techniques, such as Quantization, can create smaller, more efficient models without significant loss of performance.

Improving multilingual capabilities is another promising research direction. Extending the model's capabilities to handle multiple languages appeals to smartKYC's global customer base. To improve performance in this area, it may be useful to use transfer learning techniques to leverage knowledge from one language. This approach can increase the model's utility across regions by enabling it to understand and generate summaries in multiple languages.

Improving factual accuracy and reducing hallucinations are critical to maintaining the reliability of summaries. Incorporating recall-based methods to ground model outputs in real data can reduce hallucinations. Developing auxiliary modules that verify the actual accuracy of generated summaries before they are presented to users can increase trust and reliability.

Human-in-the-loop systems can be developed that allow users to guide the summarization process, providing input or corrections from which the model can learn. Implementing mechanisms to collect user feedback on summaries can inform ongoing model refinement. This collaborative approach can lead to continuous model refinement based on real-world usage and expert insights.

Ethics and bias considerations are central to the development of AI systems. Exploring methods to identify and reduce bias in model outputs ensures fairness

and adherence to ethical standards. Developing techniques to make the model's decision-making processes more interpretable fosters trust among users. Transparency about how the model generates summaries can help users understand and evaluate the information provided.

In addition, integration with other AI technologies can be explored to enhance the platform's capabilities. Combining summarization with other NLP tasks, such as entity recognition and sentiment analysis, provides richer insights. Using the model's capabilities to contribute to automated risk assessment tools within smartKYC can enhance the platform's functionality and value proposition.

6

Conclusion

The development of an effective summarization system for the smartKYC platform represents a significant advancement in automating compliance and due diligence processes. This thesis detailed the comprehensive approach undertaken to select and optimize a language model capable of producing accurate, consistent, and contextually relevant summaries of complex legal and financial documents.

The first phase involved a comprehensive evaluation of several state-of-the-art language models, including models such as Llama2, gemini-1.5-pro, gpt-4-turbo. Experiments and analysis revealed that the gpt-4 family models performed better in handling domain-specific content, maintaining consistency, and producing outputs that closely aligned with human-written summaries. The model's ability to accurately capture nuanced information and present it in a consistent and structured manner was critical to the smartKYC platform requirements.

Recognizing the fundamental role of prompt engineering in influencing model performance, an iterative process of prompt optimization was undertaken. In collaboration with language experts, prompts were improved to address specific challenges such as unnecessary repetition, hallucinations, wording, and formatting inconsistencies. This improvement led to significant improvements in the quality of the summaries produced, increasing accuracy, consistency, completeness, and adherence to the desired output structure. While the final summarization results showed that the gpt-4-turbo-preview model driven by optimized prompts effectively produced high-quality summaries that met the requirements of the platform, the evaluation between gpt-4o and gpt-4-turbo-preview with updated prompts is still ongoing. Quantitative evaluations

using metrics such as BLEU, METEOR, ROUGE, and BERTScore confirmed the models' better performance compared to the other evaluated models. The models achieved higher scores in all metrics and showed better agreement with human-written summaries in terms of content, linguistic expression, and semantic similarity.

Future research directions were also examined in the thesis, including the improvement of unique language models like LLaMA 3 8B and Mistral 7B. These models can be used to further increase performance while resolving data privacy concerns related to external APIs by being deployed on the AWS cloud architecture and trained using domain-specific data gathered over the past year. Custom model fine-tuning can increase operational autonomy, domain flexibility, and accuracy. In the end, it can also help the platform scale and comply with data protection laws. Developing multilingual capabilities to serve a global customer base, exploring advanced models with larger parameters, enhancing factual accuracy by integrating retrieval mechanisms, and creating human-in-the-loop systems for interactive summarisation are some of the potential research directions that have been identified. Sustaining trust and adhering to ethical norms also depend on addressing ethical concerns including decreasing prejudice and boosting model transparency.

In conclusion, this thesis has demonstrated the feasibility and effectiveness of leveraging advanced language models to automate the summarization of complex legal and financial documents on the smartKYC platform. The comprehensive approach combining model evaluation, prompt engineering, and evaluation against human metrics has targeted a robust summarization system that improves efficiency, accuracy, and user experience. The findings add valuable insights to the field of natural language processing, especially in the application of large language models to specialized domains. The continued exploration of custom model fine-tuning and the integration of human expertise holds promise for further advances, ensuring that the smartKYC platform remains at the forefront of innovation in compliance and due diligence automation.

Appendix A

Initial Prompt for Offense Aggregation Frame : "Below, delimited by triple backticks, there are snippets about watchedEntity. Each snippet is preceded by an ID of the form <idx>, where x is a number. Based on these snippets, what are the most prominent offenses associated with watchedEntity? Answer according to the response structure below such that each ""summary"" property consists of two lines (separated by a
n line break) in exactly the format below. Offense: (<Description of the Offense based on the smartKYC platform and Offense Aggregation Frame>) Stage: (<Description of the Stage based on the smartKYC platform and Offense Aggregation Frame>) Your response should be entirely in English and fit exactly the response structure below. Each line break in each ""summary"" property should be represented by
n. ""summary"": '...', ""references"": ['<idx>', '<idy>'] , ""summary"": '...', ""references"": ['<idz>', '<idt>', '<idk>'] ...] Here are the snippets: "'claimSnippets"'"

Last Prompt for Offense Aggregation Frame : "At the bottom there are snippets, delimited by triple backticks, about watchedEntity. Each snippet is preceded by an ID of the form <idx>, where x is a number. Some snippets are preceded by their publication date, in curly brackets." "Based on these snippets, provide a comprehensive summary of the bankruptcy and liquidation facts associated with watchedEntity. Answer according to the response structure above the snippets such that each ""summary"" property consists of two lines separated by
n, a line beginning with ""Type:"" and a line beginning with ""Status:"", in exactly the following format: Type: (<Description of the Type based on the smartKYC platform and Offense Aggregation Frame>)
nStatus: (<Description of the Status based on the smartKYC platform and Offense Aggregation Frame>) Do not include multiple ""summary"" properties with the same Type." "Your response should be entirely in English. Translate

any non-English text into English, and transliterate any names in non-Latin script into English. Each double quotes within each ""summary"" property should be escaped via
"". Here is the response structure: [""summary"": ""..."", ""references"": [""<idx>"", ""<idy>""] , ""summary"": ""..."", ""references"": [""<idz>"", ""<idt>"", ""<idk>""] ...] Some of the following snippets might be irrelevant to the requested summary, please do not cite them in any ""references"" property. Here are the snippets: ”’claimSnippets”’ If you are unable to summarize certain snippets as requested, end your response with a ""summary"" property which summarizes these snippets in an unstructured way."

Initial Prompt for Political Exposure Frame : "At the end of this prompt, delimited by triple backticks, there are snippets about watchedEntity. Each snippet is preceded by an ID of the form <idx>, where x is a number. Based on these snippets, provide a comprehensive summary of all familial, financial, and workplace relationships of watchedEntity with politically-exposed people (current and former politicians), political bodies, and national companies. We are not interested in relationships with prizes. Answer exactly according to the response structure above the snippets such that each ""summary"" property consists of an ""Entity"" line and a ""Relationship"" line in exactly the following format: Entity: (<Description of the Entity based on the smartKYC platform>)
nRelationship: (<Description of the Relationship based on the smartKYC platform and Political Exposure Frame>) Begin your response with the ""summary"" properties where the Entity is a person. Your response should be entirely in English. Each double quotes within each ""summary"" property should be escaped via ¨. Here is the response structure: [""summary"": ""..."", ""references"": [""<idx>"", ""<idy>""] , ""summary"": ""..."", ""references"": [""<idz>"", ""<idt>"", ""<idk>""] ...] Here are the snippets: ”’claimSnippets”’"

Last Prompt for Political Exposure Frame : "At the bottom there are snippets, delimited by triple backticks, about watchedEntity. Each snippet is preceded by an ID of the form <idx>, where x is a number. Some snippets are preceded by their publication date, in curly brackets. Based on these snippets, provide a comprehensive summary of all familial, financial, and workplace relationships of watchedEntity with politically-exposed people (current and former politicians), political bodies, and state-owned companies. We are not interested in relationships with non-politicians, non-political organisations, or prizes. Answer according to the response structure above the snippets such that each

""summary"" property consists of two lines separated by
n, a line beginning with ""Political entity:"" and a line beginning with ""Exposure:"", in exactly the following format: Political entity: (<Description of the Political entity based on the smartKYC platform and Political Exposure Frame>) nExposure: (<Description of the Political entity based on the smartKYC platform and Political Exposure Frame>) Begin your response with the ""summary"" properties where the Political entity is a person. Your response should be entirely in English. Translate any non-English text into English, and transliterate any names in non-Latin script into English. Each double quotes within each ""summary"" property should be escaped via
"". Here is the response structure: [""summary"": ""..."", ""references"": [""<idx>"", ""<idy>""], ""summary"": ""..."", ""references"": [""<idz>"", ""<idt>"", ""<idk>""] ...] Some of the following snippets might be irrelevant to the requested summary, please do not cite them in any ""references"" property. Here are the snippets: ”’claimSnippets”’ If you are unable to summarize certain snippets as requested, end your response with a ""summary"" property which summarizes these snippets in an unstructured way."

References

[1] Benjamin Yu. *Evaluating Pre-Trained Language Models on Multi-Document Summarization for Literature Reviews*. Proceedings of the Third Workshop on Scholarly Document Processing, pages 188–192, October 12–17, 2022.

[2] Cheng-Han Chiang and Hung-yi Lee. *A Closer Look into Automatic Evaluation Using Large Language Models*. October 9, 2023.

[3] Yixin Liu and others. *Benchmarking Generation and Evaluation Capabilities of Large Language Models for Instruction Controllable Summarization*. November 15, 2023.

[4] Peiwen Yuan, Shaoxiong Feng, and Yiwei Li. *BatchEval: Towards Human-like Text Evaluation*. December 31, 2023.

[5] Liyan Xu, Zhenlin Su, Mo Yu, and Jin Xu. *Identifying Factual Inconsistency in Summaries: Towards Effective Utilization of Large Language Model*. February 20, 2024.

[6] Yukyung Lee, Joonghoon Kim, and Jaehee Kim. *CheckEval: Robust Evaluation Framework using Large Language Model via Checklist*. March 27, 2024.

[7] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. *A Survey of Evaluation Metrics Used for NLG Systems*. ACM Computing Surveys, Vol. 55, No. 2, Article 26, January 2022.

[8] Chen Chen, Wei Emma Zhang, and Alireza Seyed Shakeri. *The Exploration of Knowledge-Preserving Prompts for Document Summarisation*. October 15, 2023.

[9] Louie Giray. *Prompt Engineering with ChatGPT: A Guide for Academic Writers*. June 7, 2023.

[10] Seonghyeon Ye, Hyeonbin Hwang, and Sohee Yang. *In-Context Instruction Learning*. February 28, 2023.

[11] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, and Armand Joulin. *LLaMA: Open and Efficient Foundation Language Models*. February 27, 2023.

[12] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. *INSTRUCTION TUNING WITH GPT-4*. Microsoft Research, April 6, 2023.

[13] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. *GPT4Tools: Teaching Large Language Model to Use Tools via Self-instruction*. May 30, 2023.

[14] Zijian Győző Yang and Noémi Ligeti-Nagy. *Improve Performance of Fine-tuning Language Models with Prompting*. HTE, 2015.

[15] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. February 21, 2023.

[16] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. *SummIt: Iterative Text Summarization via ChatGPT*. October 9, 2023.

[17] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. *Extractive Summarization via ChatGPT for Faithful Summary Generation*. October 9, 2023.

[18] Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. *Zero-Shot Cross-Lingual Summarization via Large Language Models*. October 24, 2023.

[19] Shaohui Zheng, Zhixu Li, Jiaan Wang, and Jianfeng Qu. *Long-Document Cross-Lingual Summarization*. December 1, 2021.

[20] Adithya Bhaskar and Alexander R. Fabbri. *Prompted Opinion Summarization with GPT-3.5*. May 23, 2023.

[21] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. *Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text*. arXiv preprint arXiv:2202.06935, 2022.

[22] Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. *Bottom-up Abstractive Summarization*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4098–4109, Brussels, Belgium, 2018. Association for Computational Linguistics.

[23] Tanya Goyal and Greg Durrett. *Annotating and Modeling Fine-grained Factuality in Summarization*. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1449–1462, 2021. Association for Computational Linguistics.

[24] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. *News Summarization and Evaluation in the Era of GPT-3*. arXiv, 2022.

[25] Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. *NEWTS: A Corpus for News Topic-focused Summarization*. Findings of the Association for Computational Linguistics: ACL 2022, pages 493–503, Dublin, Ireland, 2022. Association for Computational Linguistics.

[26] Arthur Bražinskas, Mirella Lapata, and Ivan Titov. *Few-shot Learning for Opinion Summarization*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4119–4135, 2020. Association for Computational Linguistics.

[27] Arthur Bražinskas, Mirella Lapata, and Ivan Titov. *Unsupervised Opinion Summarization as Copycat-Review Generation*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5151–5169, 2020. Association for Computational Linguistics.

[28] Benjamin Yu. *Evaluating Pre-Trained Language Models on Multi-Document Summarization for Literature Reviews*. Proceedings of the Third Workshop on Scholarly Document Processing, pages 188–192, October 12–17, 2022.

[29] Cheng-Han Chiang and Hung-yi Lee. *A Closer Look into Automatic Evaluation Using Large Language Models*. October 9, 2023.

[30] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. *Exploring Visual Prompts for Adapting Large-Scale Models*. arXiv preprint arXiv:2203.17274, 2022.

[31] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. *Language Models are Few-Shot Learners*. 2020.

[32] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. *Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks*. arXiv preprint arXiv:2211.12588, 2022.

[33] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. *Unleashing the Potential of Prompt Engineering in Large Language Models: A Comprehensive Review*. arXiv preprint arXiv:2310.14735, 2023.

[34] Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. *Contrastive Chain-of-Thought Prompting*. arXiv preprint arXiv:2311.09277, 2023.

[35] Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. *Rephrase and Respond: Let Large Language Models Ask Better Questions for Themselves*. arXiv preprint arXiv:2311.04205, 2023.

[36] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. *Chain-of-Verification Reduces Hallucination in Large Language Models*. arXiv preprint arXiv:2309.11495, 2023.

[37] Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. *Active Prompting with Chain-of-Thought for Large Language Models*. arXiv preprint arXiv:2302.12246, 2023.

[38] Yixin Liu and others. *Benchmarking Generation and Evaluation Capabilities of Large Language Models for Instruction Controllable Summarization*. November 15, 2023.

[39] Peiwen Yuan, Shaoxiong Feng, and Yiwei Li. *BatchEval: Towards Human-like Text Evaluation*. December 31, 2023.

[40] Liyan Xu, Zhenlin Su, Mo Yu, and Jin Xu. *Identifying Factual Inconsistency in Summaries: Towards Effective Utilization of Large Language Model*. February 20, 2024.

[41] Yukyung Lee, Joonghoon Kim, and Jaehee Kim. *CheckEval: Robust Evaluation Framework using Large Language Model via Checklist*. March 27, 2024.

[42] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. *A Survey of Evaluation Metrics Used for NLG Systems*. ACM Computing Surveys, Vol. 55, No. 2, Article 26, January 2022.

[43] Chen Chen, Wei Emma Zhang, and Alireza Seyed Shakeri. *The Exploration of Knowledge-Preserving Prompts for Document Summarisation*. October 15, 2023.

[44] Louie Giray. *Prompt Engineering with ChatGPT: A Guide for Academic Writers*. June 7, 2023.

[45] Seonghyeon Ye, Hyeonbin Hwang, and Sohee Yang. *In-Context Instruction Learning*. February 28, 2023.

[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, and Armand Joulin. *LLaMA: Open and Efficient Foundation Language Models*. February 27, 2023.

[47] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. *INSTRUCTION TUNING WITH GPT-4*. Microsoft Research, April 6, 2023.

[48] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. *GPT4Tools: Teaching Large Language Model to Use Tools via Self-instruction*. May 30, 2023.

[49] Zijian Győző Yang and Noémi Ligeti-Nagy. *Improve Performance of Fine-tuning Language Models with Prompting*. HTE, August 26, 2015.

[50] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. February 21, 2023.

[51] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. *SummIt: Iterative Text Summarization via ChatGPT*. October 9, 2023.

[52] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. *Extractive Summarization via ChatGPT for Faithful Summary Generation*. October 9, 2023.

[53] Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. *Zero-Shot Cross-Lingual Summarization via Large Language Models*. October 24, 2023.

[54] Shaohui Zheng, Zhixu Li, Jiaan Wang, and Jianfeng Qu. *Long-Document Cross-Lingual Summarization*. December 1, 2021.

[55] Adithya Bhaskar and Alexander R. Fabbri. *Prompted Opinion Summarization with GPT-3.5*. May 23, 2023.

[56] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. *Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text*. arXiv preprint arXiv:2202.06935, 2022.

[57] Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. *Bottom-up Abstractive Summarization*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4098–4109, Brussels, Belgium, 2018. Association for Computational Linguistics.

[58] Tanya Goyal and Greg Durrett. *Annotating and Modeling Fine-grained Factuality in Summarization*. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1449–1462, 2021. Association for Computational Linguistics.

[59] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. *News Summarization and Evaluation in the Era of GPT-3*. arXiv, 2022.

[60] Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. *NEWTS: A Corpus for News Topic-focused Summarization*. Findings of the Association for Computational Linguistics: ACL 2022, pages 493–503, Dublin, Ireland, 2022. Association for Computational Linguistics.

[61] Arthur Bražinskas, Mirella Lapata, and Ivan Titov. *Few-shot Learning for Opinion Summarization*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4119–4135, 2020. Association for Computational Linguistics.

[62] Arthur Bražinskas, Mirella Lapata, and Ivan Titov. *Unsupervised Opinion Summarization as Copycat-Review Generation*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5151–5169, 2020. Association for Computational Linguistics.

[63] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. *BLEU: a Method for Automatic Evaluation of Machine Translation*.Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

[64] Satanjeev Banerjee and Alon Lavie. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, June 2005 Association for Computational Linguistics.

[65] Chin-Yew Lin. *ROUGE: A Package for Automatic Evaluation of Summaries*. Conference: In Proceedings of the Workshop on Text Summarization Branches Out 2004

[66] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi. *BERTSCORE: EVALUATING TEXT GENERATION WITH BERT*. arXiv:1904.09675v3 [cs.CL] 24 Feb 2020

[67] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N.,Polosukhin, I. *Attention is all you need. In Advances in neural information processing systems* .(pp. 5998-6008) (2017).