Enhancing the Accuracy of Abstractive and Extractive Summarization of Patient Discharge Reports by using Transfer models

Gaurav Gupta
x22212311@student.ncirl.ie
National College of Ireland

Abstract

We will conduct our research on extractive and abstractive text summarization in the medical field. The accuracy of medical report summaries has been fairly limited up until now, but since it will save doctors and patients time, it is time for improvements. The National Institutes of Health states that it takes 60 hours on average to handle a patient's discharge report and also the ratio of doctors to patients in developing and poor nations is approximately 0.05 %. Sentence-to-sentence models and encode-decoders have been the focus of much previous research on this problem, but more recent work has shown that abstractive summarization techniques like T5, DistilBART, and Pegasus, as well as extractive summarization techniques like BERT-SUM and XLNet, can produce better results. Analyze the generated text by ROUGE and BERT score and evaluate on the basis of relevance, coherence, and fluency of the generated text.

1 Introduction

The use of natural language processing in the health sector has transformed patient care and the medical industry in recent years. Understanding and interpreting human text is the focus of NLP. At the moment, it is essential to medical literature, insurance summaries, and electronic health records. However, there hasn't been much progress in the area of condensing lengthy

discharge summaries up to this point. This is because most of the transfer learning models for summarisation have been produced recently as well as the confidentiality of patient information, data privacy, and ethical requirements all place significant restrictions on the access of healthcare data.

This paper addresses the research question of **"How can transfer models enhance the accuracy of text summarisation of patient discharge reports?"**

1.1 Objective

When considering the roles of physicians and patients, I believe that this is a significant and persistent issue, since it takes hours for physicians to comprehend and analyze the reports. At this point, physicians could be employed elsewhere to benefit society more broadly rather than thoroughly reading reports. If the report is explained in simple, easy-to-understand language, it is also highly beneficial for the patients. This problem hasn't been very accurate up to now in terms of relevance, consistency, clarity, coherence (logical flow), and fluency. Therefore, the accuracy can be improved and enhanced with this study.

1.2 Motivation

Earlier research papers have either concerned abstractive or extractive summarisation Earlier research papers have focused on extractive summarisation where information is selected and stripped off of documents or the papers have failed to create new documents out of the source documents. On the other hand, some scholars focus their attention on Trivial or abstractive summarization which entail text synthesis in order to make sense of it and improve the coherence. However, the field is extremely tough because it is hard for machines to interpret them semantically from original texts. Thus, with the aid of topic modeling, we shall conduct our research in both the extractive and abstractive fields. When it comes to Natural Language Processing research, data modeling is vital, but so is figuring out how to evaluate our model. A model's evaluation in previous works is based on its ROUGE score. Overall readability, logical flow, and summary fluency are not taken into account. Thus, we will explore additional performance metrics in this work, including the BERT score, ROUGE variations, and BLEU score—all of which are utilized by Facebook and Google for text evaluation.

2 Related Works

In section 2, we will look into all the progress happening in this area, approach all the research papers, compare their results from each other and how we can rectify their results and come up with our new research model.

2.1 Extractive Text Summarisation

The foundation of all earlier research was extractive text summarization. Thus, we will examine prior research and methods related to extractive summarization in this part. DeepSumm, as proposed by Ghadimi and Beigy (2023) and Verma et al. (2022), joins the text's structural and semantic context with the summary that is to be produced. DeepSumm uses this technology of word embeddings and topic distributions to produce document sentences that are applicable to both content and context. Contextual sentence selection techniques employ sequence-to-sequence networks, which can assign scores based on subject and content. Additionally, the study offers a fresh strategy for SNS that emphasizes innovation as well as non-redundant and varied summaries. In order to maximize the level of supervision for extractive summarizing, future work will focus on abstracting features for abstractive summarization and evaluating unsupervised techniques based on topics. Zhang et al. (2024) and Xie et al. (2022) develop the SGCSum multi-document summarizing technique and assess its effectiveness using a healthcare dataset. The BERT pre-trained language model uses graph theory and a graph convolutional network for feature learning to represent the text. Compared to other summarizing methods, this produces superior results, and the amount of convolution pooling layers affects performance. However, there are benefits to abstractive summary over extractive summarization, such as the capacity to provide summaries that are more condensed and logical and the ability to convey a text's overall meaning. I'll make an effort to view things similarly.

2.2 Abstractive Text Summarisation

The researcher has selected the medical chat set (33,699 doctor-patient chats) from Searle et al. (2023) and Aaron M Silver et al. (2022). The researcher used three different strategies: LIME for task highlighting, TF-IDF scores to carry out the text's critical units (although TF-IDF lacks contextual informa-

tion), and LSTM, which is for sequential modeling and can learn both long- and short-term dependencies while producing the best results. Utilize ROC-AUC, PR-AUC, and precision-recall curves to assess the performance. In the future, employing the BERT approach. In addition, it proposes increasing testing models and transformers attention methods.

Two datasets were employed by Ozyegen et al. (2022) and Torres-Parejo Úrsula et al. (2021): KCH, which had 34,179 unique documents, and MIMIC-III, which had 1,441,109 unique documents. They used Text-Rank and BI-LSTM for both extractive and abstractive summarization, and BERT and T5 for abstractive summarization. It suggests employing BERT score in addition to ROUGE, utilizing MEDCAT and other guidance signals. The efficiency of several summarization methods in the medical field is conducted and compared by the researchers. based on LSTM rankings and semantic contextual embedding analysis of extraction models. Compared to the other assessed models, BART, which is built on PubMed pre-trained transformer abstractive models, showed a noticeably higher capacity for case summarization.

Additionally, utilizing pre-trained publications models to emphasize abstractive summaries, medical signals generated with MedCAT's assistance increased their quality. The ensemble summation methods tested every option, but the improvements were negligible. The assessment of the guiding signals was crucial in obtaining summaries that were applicable to clinical settings. For additional effort, such as creating training signals that are more precise and investigating various assessment methods. According to Joshi et al. (2023) and Issam et al. (2021), information extraction—which includes decision-making and the extraction of information—is crucial in the healthcare industry. One key way to combine dictionary-based, rule-based, and deep learning techniques is the rule-based model. These methods address the problems of multilingual support, data annotation, and domain-specific flexibility. The outcome is enhanced by methods such as domain-specific BERT, CRF and BiLSTM combination models, and dynamic rule creation. However, this work has issues with document summarizing and medical data analysis in these systems. The challenge of accurately extracting text from a large number of medical texts when summarizing papers using NER remains. Transfer learning will be used in future work to condensibly retrieve knowledge from many sources. Van Veen et al. (2023) concentrate on the GPT 3.5 and GPT 4 document-level performance of LLM models in simulating the MIMIC-III dataset. Among other models, the study assesses the FLAN-T5,

GPT-3.5, and GPT-4 models.

All in all, the content in GPT 4 is more detailed, precise, and concise compared to that of GLTR and GPT 3. Thus, BERTScore is used for evaluating the text for correctness along with the MEDCON score while the text quality is assessed with the BLEU score. As for the future work, incorporating the transformer learning and sequence to sequence models into the process to enhance the quality and capacity of the generated summaries is recommended. Bisen W et al. (2024) while they state that a brief description of the specifics in questions within the literature text needs to be created. Thus, the process involves pre-preprocessing of data that let incorporates textual information into training neural networks easily. Unlike other kind of techniques where entire phrases, sub-phrases, or complete sentences are translated from the source material, abstractive summarization summary resembles a kind of writing that would be done by an actual human being.

When it comes to content summarization, there are two primary categories of methodologies to consider: On the basis of structure we can approach it with methods on structure ad on the basis of meaning there are methods on meaning. Consequently, the trees are exploited by the structural approaches to manage and analyze the Textual Relations within phrases. In contrast, semantically based approaches use videos and graphics to expand the analysis of the text's meaning even further to an extent. The models are somewhat less accurate in the issues of ambiguity and context. Hence to enhance the context and comprehension of the text for both the physicians and patients, there should be exploration of both the extraction as well as summarization techniques in the subsequent study.

2.3 Evaluation Techniques

In this section, I'll go over evaluation strategies that haven't been applied before in this field or for this particular issue but that can be applied to improve performance, assessment, and the text's general context. As pointed by Mohan et al. (2024) and G. Bharathi Mohan et al. (2023), the evaluated metrics include Coherence, Consistency, Fluency and Relevance are measured with ROUGE-SEM ,SummEval and DialSummEval. Also incriminated concerning the results is the fact that distinct levels of connection between the said dimensions and indicators are observed. The analysis is provided by BERTScore based on human evaluations of continuity and readability. Nevertheless, there is not a great deal of statistical evidence for its relevance

and the objectivity of the consistency assessments provided by people. Better still, ROUGE-2 was established to indicate results closer to the human judgments than METEOR or BLEU. With the help of several attributes, the use of metrics related to semantic embedding, namely BERTScore, can be considered as more effective in all four categories. It is observed after comparing the ROUGE-SEM variants with the ROUGE variants on SummEval and DialSummEval datasets that the proposed ROUGE-SEM performs better in several aspects. Pecifically, our hypothesis about the difference in the effectiveness of ROUGE-SEM displays with regard to consistency and coherency as well as their relevance is confirmed,TO Bastin et al. 's report which proves that the use of the ROUGE-SEM for purpose of the automatic method of the summarization evaluation is more effective in contrast with the standard ROUGE as well as embedding techniques. According to Landolsi et al. (2023) and Shi J et al. (2022), the T5 models performed well in the domain of lexical and syntactic coherence when they were evaluated using a variety of metrics, including ROUGE, METEOR, BERTScore, Cosine Similarity degree, and BLEU. Semantic continuity is effectively maintained by PEGASUS and BART. Keswani et al. (2024) and Touvron, H et al. (2023) Retrieval-Augmented Generation (RAG) systems analyze both the combined Score and the retrieval and generation separately in order to assess performance. The researcher presents new metrics in RAG to compute answer relevancy, fidelity, recall, and context relevance.

These metrics aid in producing responses that appropriately address the topic and its context. Apart from using n-grams to measure recall, the ROUGE score provides machine-generated summaries in comparison to reference summaries. Large language models, vector similarity, and abstract clustering methods are used by the researcher. This approach of summarizing guarantees multilingual capabilities, contextual accuracy, and comprehensive, domain-specific summaries.

2.4 Gap Analysis

It is clear from the preceding conversation that there are still some unsolved questions and that this field needs more research. In terms of approach and assessment, there is still a gap, all the previous model is either providing abstractive text summarisation or extractive text summarisation but what if we required both of a text, and also enhancing the accuracy of each one. I'll examine how applying transfer learning can improve accuracy in the fol-

lowing section. I will use five different transfer learning algorithms which are T5, DistilBART and Pegasus for abstractive summarisation, BERTSUM and XLNet for extractive summarisation. In order to quantify the performance, most of the paper using ROC, precision and recall but that is not correct technique to evaluate text as our aim to evaluate quality and relevance of the generated text so I will also investigate new evaluation metrics that have not yet been applied to this problem, such as BLEU, ROUGE-1, ROUGE-2 and ROUGE-L.

3 Methodology

The research project demonstrates text summarisation of patient discharge reports to enhance the context, relevance, recall and relevancy. In this paper, we will use the CRISP-DM methodology which has the following steps : Business understanding, Data understanding, Data preparation, Modeling, Evaluation.

3.1 Business understanding

Collecting requirements and business objectives is the first stage. NLP can enhance the quality of patient discharge reports, which is beneficial for those involved in the healthcare industry.If medical professionals are able to accurately generate discharge report summaries, it can save them a great deal of time. This could facilitate more efficient use of healthcare resources and assist allocate them in an orderly fashion. Additionally, patients can better comprehend their diseases by receiving clear and intelligible information. Through the automation of discharge summaries, this can also lower healthcare costs (human costs). Additionally, since this paper can be used as a baseline model for future research and development, these insights may also be helpful in that regard.

3.2 Data Gathering and Understanding

The literature review demonstrates the availability of several datasets. The largest dataset, however, is MIMIC-III (Medical Information Mart for Intensive Care III), which was gathered between 2001 and 2012 at Shabbat Israel

Deaconess Medical Center. It is a sizable, varied, and openly accessible collection of different patient ICU records. 1,12,000 patient reports with an average length of 709 tokens make up the dataset. This dataset contains information on survival rates, length of patient stay, imaging reports, discharge reports, and laboratory test results, among other things. We will focus on discharge reports of patients in this paper.

3.3 Data Preprocessing

The next step after choosing the dataset is to analyze the data and reduce the noise in the data accordingly. The dataset is visualized by checking the length of each text, the most common words in the text. In addition, handling abbreviations, filling missing text with empty string, HTML elements, punctuation, and special characters must be eliminated during this process. The data must next be tokenized, which entails turning text into tokens and then eliminating stopwords from tokenized words like "is" and "the." The next step involves vectorization to transform text data into numerical representation and stemming or lemmatization to lower dimensionality.

3.4 Modeling

In this section, we will look into algorithms for text summarizing pre-trained models which have higher performance,accuracy and require very less time to train the model.

- **PEGASUS**: PEGASUS (Figure 1) is a Transformer model that is conditioned via the encoder-decoder mechanism for abstractive text summarization. Its peculiar pre-training objective is to mask complete sentences and then try to predict them, and it is evident that this correlates well with the summarization task. This approach has one significant advantage and therefore better performance when dealing with the generation of short and coherent summaries from long and large texts, such as medical ones. Despite the fact that training from scratch is computationally expensive, fine-tuning PEGASUS from scratch on domain-specialized data is quite doable, and will result in highly effective measures of summarization. Since PEGASUS is part of the Hugging Face Transformers library, it can easily be used and

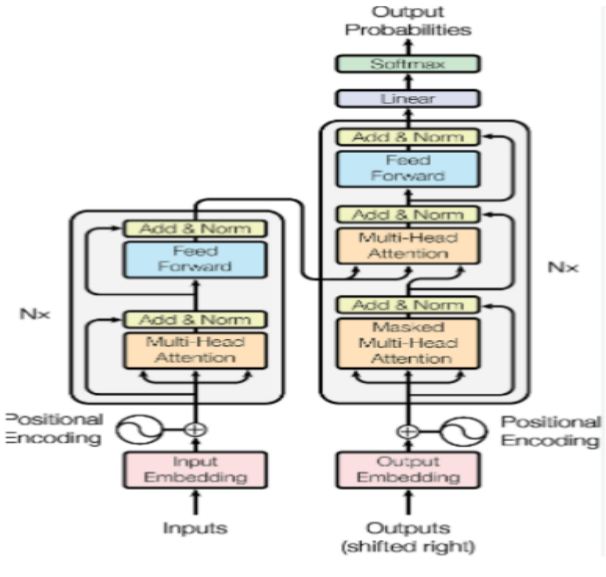

Figure 1: Pegasus Architecture

fine-tuned for the task, thus adding more versatility to its functionality for medical text summarization.

- **T5 (Text-To-Text Transfer Transformer)**: T5 (Figure 2) is another general-purpose Transformer model that reformulates all the NLP challenges as a text-to-text issue. It is capable of performing finely throughout a variety of functions because of this unification, and summarization just happens to be among them. Because of its adaptability and relatively study architectural design, it is well-positioned to deal with medical text summarization especially when it is fine-tuned on certain datasets that are within the medical genre. Nevertheless, T5 can be costly when it comes to computational resources; however, T5-small and T5-Base are more realistic, a promising direction is used in papers. Due to the availability of many pre-trained models and the substantial amount of support offered by the Hugging Face library, T5 is suitable for summarizing medical texts.

- **BERTSum**: BERTSum (Figure 3) is one of the forms of applying BERT that is focused on the summarization process, more specifically extractive one. In addition to that, BERTSum introduces layers

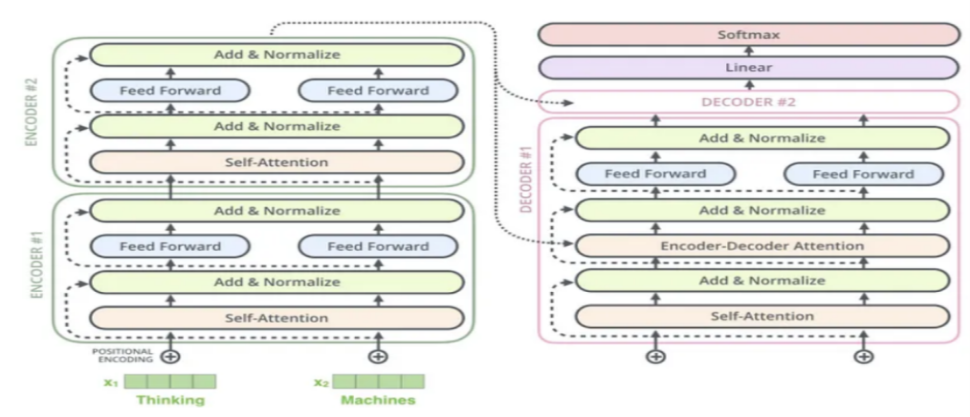

Figure 2: T5 Architecture

for summarization that utilizes BERT's masked language model pre-training to select key sentences and produce summaries. This means that this model is well suited for use in extractive summarization hence can be used in situations where key sentences from Medical texts need to be highlighted. However, it could need some changes for the abstractive summarization type as well. Compared to other full encoder-decoder models, BERTSum is more resource-friendly, which is why it is more efficient for extractive tasks compared to, for instance, PEGASUS or T5 but may lack the fine details of abstract summarization.

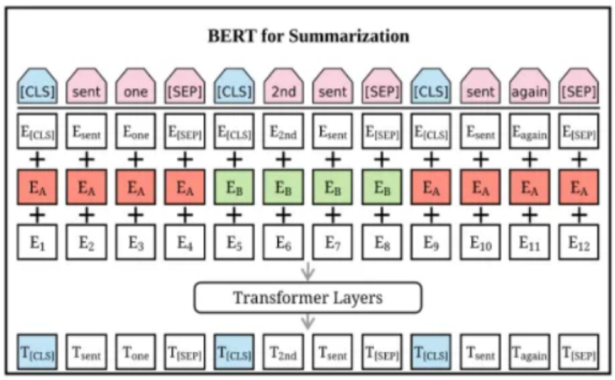

Figure 3: BERTSum Architecture

- **DistilBART**: By the merit of its mark down DistilBART (Figure 4) is much smaller and faster than BART; however, it retains almost 95% of the capacity of BART in generation of text and summarization. This model also balances the power and speed hence it becomes relevant to summarization tasks that require little power. DistilBART can then directly be fine-tuned for medical text summarization, which then makes a lot of sense and offers coherent, though less-sized summaries with slower inference time as a trade-off. The fact that it is included in the Hugging Face Transformers library also ensures its functionality and access; therefore, it is possible to state that DistilBART can be considered a useful tool to apply for medical text summarizing.

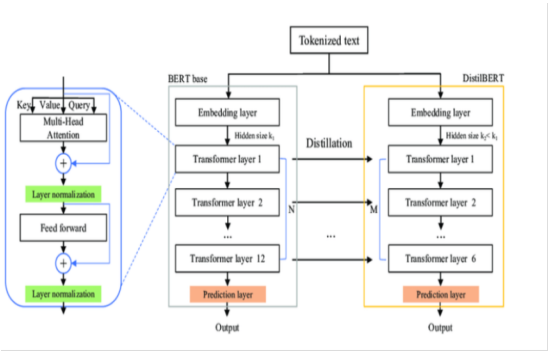

Figure 4: DistilBART Architecture

- **XLNet**: XLNet (figure 5) is a permutation-based Transformer model that is comfortable with bidirectional contextual learning without leaving out any permutation of the word order during training. Although unsuitable for permutation invariant outputs like summarization, due to its training method, XLNet is a superior language model that can easily be fine-tuned for NLP in general, and thus, for summarization in particular. Regarding the summarization, it is also possible to train XLNet to generate the summaries however, it may not be as accurate as PEGASUS or T5, both of which are designed for the medical text summarization. However, as far as its performance is concerned, due to its capability to capture long-distance contextual relationships that boosts its overall power, one could post XLNet as a viable one if and

only if it will be trained with medical data pertinent to the specific relevant domain.

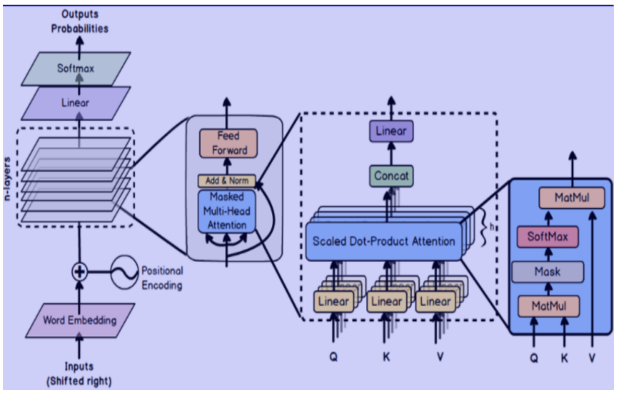

Figure 5: XLNet Architecture

3.5 Evaluation

Assessing summarization is crucial for ensuring the effectiveness and caliber of the summary. Coherence, readability, efficacy, and relevancy of the original material are factors in determining a summarization's quality. It also aids in comprehending how various algorithms perform. The following evaluation criteria will be applied in order to assess our model's performance:

- **ROUGE and Its Conversions**: ROUGE is helpful in guaranteeing the quality of the synopsis. It assesses summaries that are comparable to the idea of what human-written summaries are like. We will make use of ROUGE-1, ROUGE-2, and ROUGE-L for this paper.
  - **ROUGE-1**: Measures basic overlapping of individual words and usually provides rather high scores.
  - **ROUGE-2**: Coverage of bigrams and usually gives a lower score as compared to ROUGE-1.
  - **ROUGE-L**: Measures the longest common subsequence which gives an idea about the structure of the sentences.

$$\text{ROUGE-1} = \frac{\sum_{S \in \text{Reference Summaries}} \sum_{\text{gram}_1 \in S} \min\left(\text{Count}_1(\text{gram}_1, \text{Candidate Summary}), \text{Count}_1(\text{gram}_1, S)\right)}{\sum_{S \in \text{Reference Summaries}} \sum_{\text{gram}_1 \in S} \text{Count}_1(\text{gram}_1, S)}$$

$$\text{ROUGE-2} = \frac{\sum_{S \in \text{Reference Summaries}} \sum_{\text{gram}_2 \in S} \min\left(\text{Count}_2(\text{gram}_2, \text{Candidate Summary}), \text{Count}_2(\text{gram}_2, S)\right)}{\sum_{S \in \text{Reference Summaries}} \sum_{\text{gram}_2 \in S} \text{Count}_2(\text{gram}_2, S)}$$

$$\text{ROUGE-L} = \frac{\sum_{S \in \text{Reference Summaries}} \text{LCS}(\text{Candidate Summary}, S)}{\sum_{S \in \text{Reference Summaries}} \text{Length}(S)}$$

- **BLEU and its Variations**: BLEU is crucial for assessing text quality based on how similar the text is to reference translations. It emphasizes precision over n-grams.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r \end{cases}$$

3.6 Deployment

This is mainly the future work of the project. Create a web application using Flask, containerization of an application using Docker and deploy it over AWS, Azure or GCP.

4 Design Specification

The requirement for the design specification for this work that serves as the foundation for the subsequent work is to develop a text summarisation system that is specific to the patient discharge reports and the key goal of enhancement to the extent of contextual, pertinent, recall-oriented, as well as explicitly clear information. This system is applicable to anybody who forms part of the health maintaining/planning/caring institution or individually, medical personnel, patient, health care manager, or researcher because it among others, cuts down on time that is required to perform things, help develop resource utilization, expound patient outlines while making general health care affordable since majority of them processes will be automated. This makes it possible for this project to employ among other databases the MIMIC-III dataset and more specifically patient discharge reports. It includes text cleaning, handling an abbreviation, transforming the text into tokens, removing stop words and Vectorizing are the basic steps of preprocessing. Such models will include PEGASUS, T5, BERTSum, DistilBART, and XLNet, and the system will adjust it based on the type of summarization; abstractive or extractive. Thus, to evaluate the performance of the proposed models ROUGE-1, ROUGE-2, ROUGE-L, as well as BLEU coefficients will be used. The deployment process is to operate the created web application based on Flask and Docker and start it on Amazon Web Service, Microsoft Azure or Google Cloud Platform. Future work related directions are connected with increasing the size and enhancing the quality of the obtained dataset, improving the presented model, as well as enhancing the pipeline of the model usage with the focus on its effectiveness and the amount of work that can be effectively processed. This can therefore be described as a representational prescription of the structural design (Figure 6) of the system that will enable the achievement of the laid down objectives in the system

5 Implementation

The implementation phase of our research project involves translating the design architecture and methodology into a working system.

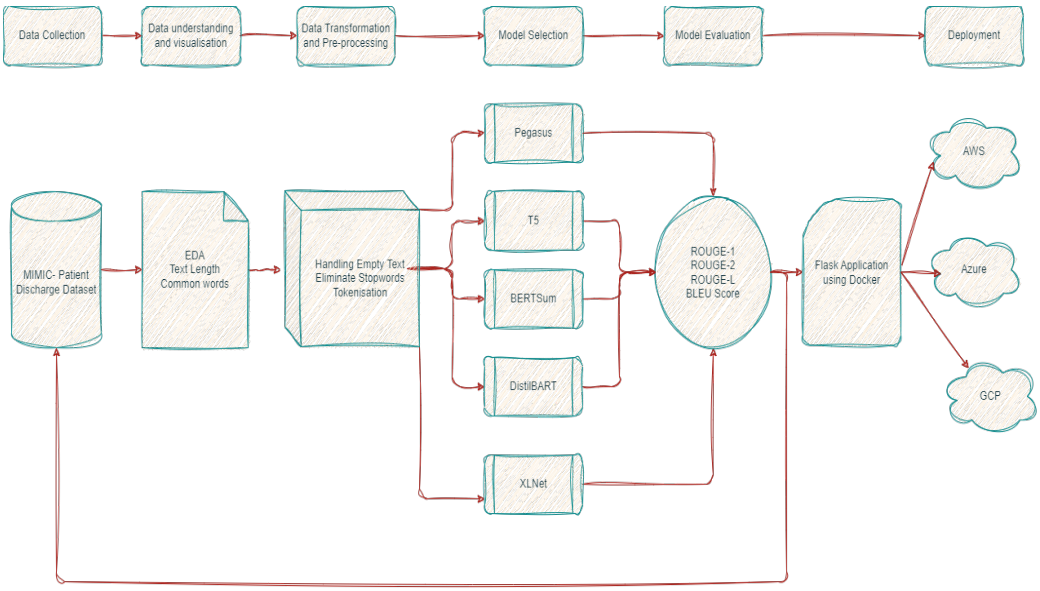

Figure 6: Flow chart of the Proposed model

5.1 Data Loading and Exploring

Data is stored as notevents.csv file, so load the dataset into google collab for further analysis and implementation. The Dataset can be described as it has 11 columns but we have to deal with only the Text column as it only has relevant text. Figure shows the average length of characters in the TEXT column which is around 8000-1000 characters as in Figure 7. And the most common 20 words in the text is Table 1

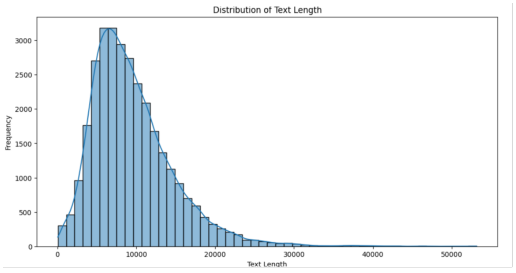

Figure 7: Distribution of Text Length

Word	Frequency
*	5,872,357
.	2,842,846
,	1,841,360
::	1,645,210
]	1,340,280
the	1,121,500
)	1,053,864
(	1,028,697
and	959,773
of	810,882
to	793,452
was	791,039
with	569,373
a	534,296
on	517,851
in	425,682
for	402,726
no	338,645
patient	332,829

Table 1: Most Common Words and Their Frequencies

5.2 Data Preprocessing

As seen in fig 7, most common words should not be required while summarizing data so it is important to remove dataset. So, it is important to do Tokenization, stopword removal, and vectorization.

5.3 Data Preparation

We have text data but we do not have summarized text according to this. So, we required corresponding summarized text associated with all the TEXT data. As data is so big, it is not possible to do it for all data points so we randomly select 237 data points and produce summarized text on my own. It will work fine as I am using transfer learning models which are already trained on millions of text data. I have also used Inter-annotator agreement to establish the degree of overlap between different human assessors of randomly

selected 10 summaries by my friend for summary quality assessment.

admission date discharge date service addendum radiologic studies radiologic studies also included chest ct con
admission date discharge date date birth sex f service history present illness patient yearold female complex mec
admission date discharge date service icu history present illness patient yearold female admitted mental status c
admission date discharge date service ccu addendum discharge medications enalapril po bid lasix po qd digoxin
admission date discharge date date birth sex f service addendum neurological patient mri eeg evaluate neurologi
admission date death date service medicinedoctor last name history present illness patient yearold male history e

Figure 8: Manual summarised Text

5.4 Data Splitting

To develop and test the accuracy of our proposed model, we partitioned the data set into 80:20 training and testing data sets. It makes sure the model is endowed with adequate amounts of data (80%) to feed on, without having access to the other part, which defines its performance (20%).

5.5 Model Selection and Training

Several modern models, namely PEGASUS, T5, BERTSum, DistilBART, and XLNet were considered for their use in the work. For the implementation, we chose PEGASUS, DistilBART and T5 for summarization as both of them work for abstractive summarization, and for extractive summarization we chose BERTSum and DistilBART. The figure 9 shows how after applying the T5 model, our summarized text looks like.

ABDOMINAL CT: Head CT showed no intracranial hemorrhage or mass effect. a chest CT confirmed cavitary
the patient is a 70-year-old female with a complex medical history. she was admitted after a cardiac arrest on
the patient is an 84 year-old woman admitted with inflammatory bowel disease. she was admitted with a histor
the patient should have potassium followed in a couple of days and monitored closely and her potassium dose
the patient had an MRI and EEG to evaluate neurologic status. the MRI showed diffuse encephalopathy and t
the patient is a 78-year-old male with a history of encephalitis, oral cancer. the patient had shortness of breath

Figure 9: Summarised Data after T5 model

6 Evaluation

In this section we will evaluate and interpret the model. In order to evaluate the different models, we required different metrics to compare the original text to generated text in terms of multiple parameters. So, we used two most popular text summary evaluation metrics: ROUGE and BLEU score. In its score summative, ROUGE essentially emphasizes recall, that is, how much of the reference text is included in the generated one. It is commonly used for summarization and there are some modifications to it called variations upon the basic algorithm. BLEU Score focuses, in fact, on the specificity, aiming to measure the extent to which the produced text matches the provided reference text. This is often used in the process of translation from one language to another. To prevent too short translation, BLEU employs a penalty to penalize short translations to avoid extreme precision accompanied by less wordiness.

6.1 Experiment 1 - PEGASUS

The Pegasus model gives the histogram of the summary length where most of the summaries are within 200 words and some go up to 1000 words as in Figure 10. The word cloud of the keywords in the summaries in Figure 11; the largest words are patient, history, hospital, and emergency medicine. The table displays the evaluation scores of the Pegasus model; thus, the ROUGE-1 score is 0. 608, ROUGE-2 at 0169, ROUGE-L at 0. 495, and for BLEU 0.486. The above metrics imply that the model works fairly good and brings a rather good amount of relevant information in the summary.

PEGASUS	**Average Score**
ROUGE-1	0.608
ROUGE-2	0.169
ROUGE-L	0.495
BLEU	0.621

6.2 Experiment 2 - T5

The table shows the ROUGE-1 score of 0.618, ROUGE-2 at 0.356, ROUGE-L at 0. 540, and BLEU at 0.628, which made me realize that this model touches on vital information that needs to be extracted and stored. The histogram

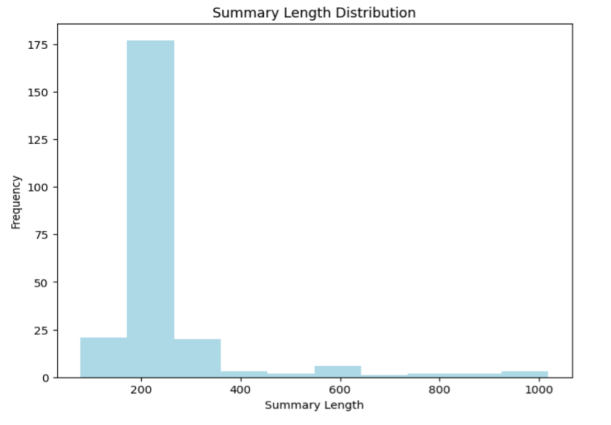

Figure 10: Summary Length by Pegasus

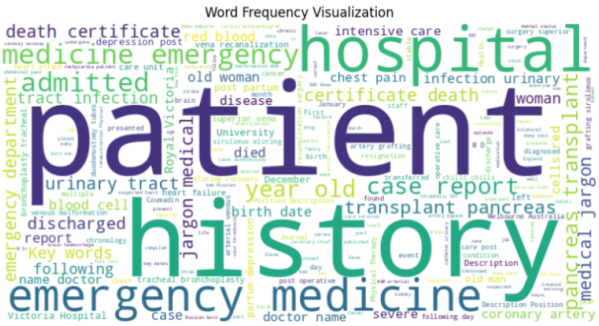

Figure 11: Word Frequency Chart

of the summary length as in Figure 12, positioned that the majority of the summaries are ranging from 150 to 350 word length, with the frequency ranging between 200 and 250 word length summaries. The word cloud in Figure 13 recaps the most often used words in the summaries with 'patient, history, year old, admitted' taking the largest portion.

T5	Average Score
ROUGE-1	0.618
ROUGE-2	0.356
ROUGE-L	0.540
BLEU	0.628

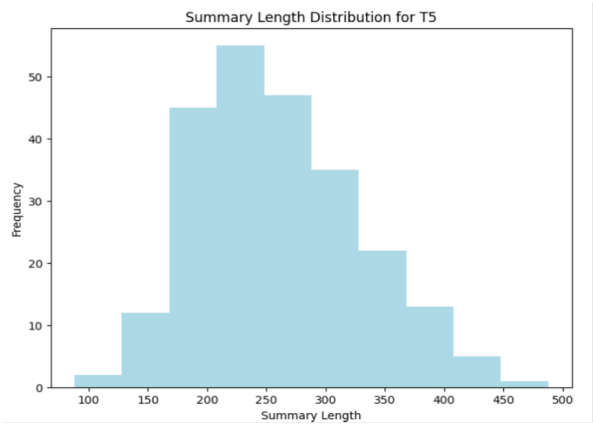

Figure 12: Summary Length by T5

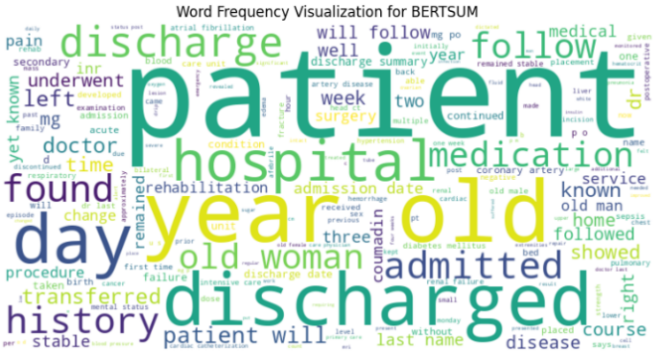

Figure 13: Word Frequency Chart

6.3 Experiment 3 - BERTSUM

The BERTSUM model for summarizing medical text indicates the evaluation metrics, and the ROUGE-1 is at 0.595, ROUGE-2 at 0.267, ROUGE-L at 0.497, and BLEU at 0.606. These scores indicate a favorable outcome of the system performance in achieving a clear and coherent summary of the abstracts as well as their relevance. The histogram in Figure 14 represents the lengths of summaries; the majority of which have between 150 and 350 words with a high of 200 words. The word cloud in Figure 15 shows the review summaries and the most frequently used terms are ‘patient,’ ‘year old,’ ‘discharged,’ ‘hospital,’ and ‘admitted.’ This visualization proves the

hypothesis that the BERTSUM model is able to capture Relevant Medical Details & Important Patient Information as equally as PEGASUS and T5 models.

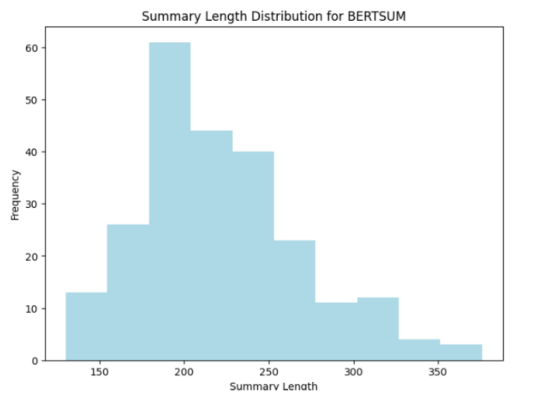

Figure 14: Summary Length by BERTSUM

BERTSUM	Average Score
ROUGE-1	0.599
ROUGE-2	0.267
ROUGE-L	0.497
BLEU	0.606

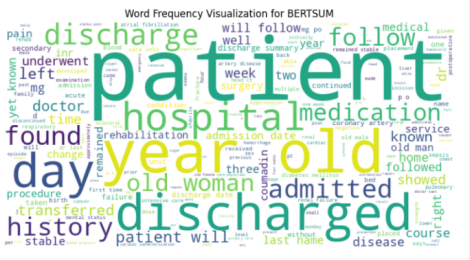

Figure 15: Word Frequency Chart

6.4 Experiment 4 - DistilBART

The DistilBART model's performance metrics and visualizations on a text summarizing job are shown in the image. The model's scores are displayed in the table: BLEU at 0.600, ROUGE-1 at 0.671, ROUGE-2 at 0.416, and ROUGE-L at 0.608. These results indicate high-quality summaries. The histogram as in Figure 16 shows that 200–300 words is the average length for summaries. The word cloud in Figure 17 illustrates often used terms like "patient," "discharged," "hospital," and "admitted," indicating that the model successfully extracts important medical data.

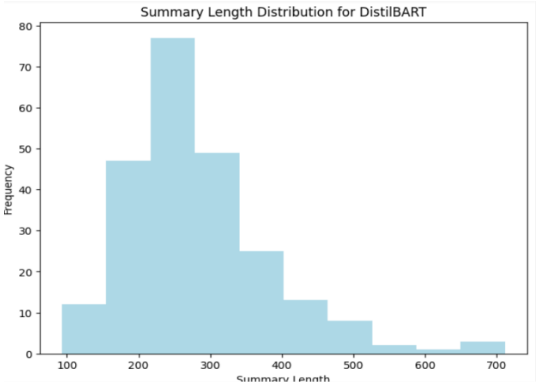

Figure 16: Summary Length by DistilBART

DistilBART	Average Score
ROUGE-1	0.671
ROUGE-2	0.418
ROUGE-L	0.608
BLEU	0.600

6.5 Experiment 5 - XLNet

The picture shows the XLNet model's performance metrics and visualizations during a text summarizing assignment. The model's scores are displayed in the table as follows: BLEU at 0.634, ROUGE-1 at 0.614, ROUGE-2 at 0.519, and ROUGE-L at 0.570. According to the histogram as in Figure 18,

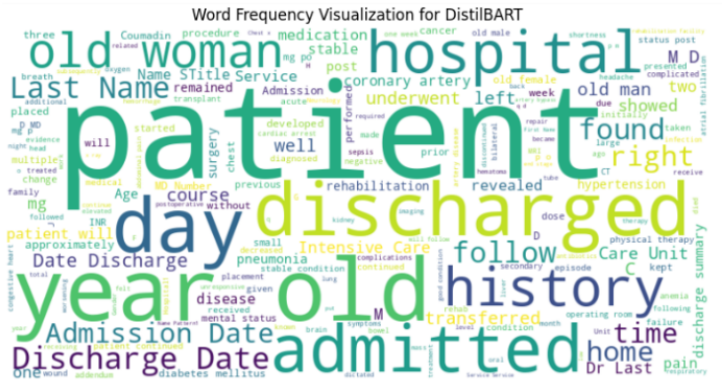

Figure 17: Word Frequency Chart

the majority of summaries have a word count of 300–600. The word cloud in Figure 19 illustrates commonly used terms like "patient," "last name," "admission date," and "discharge date," suggesting that the model prioritizes gathering pertinent data about patients and hospitals.

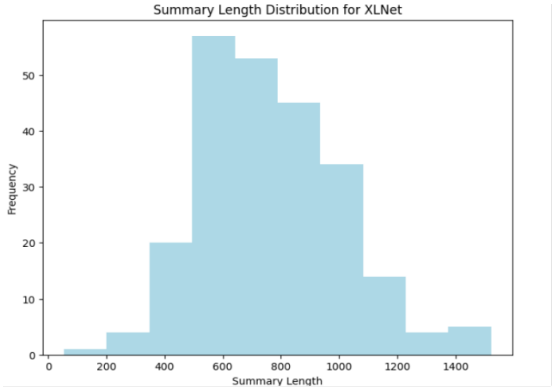

Figure 18: Summary Length by XLNet

XLNet	Average Score
ROUGE-1	0.614
ROUGE-2	0.519
ROUGE-L	0.570
BLEU	0.634

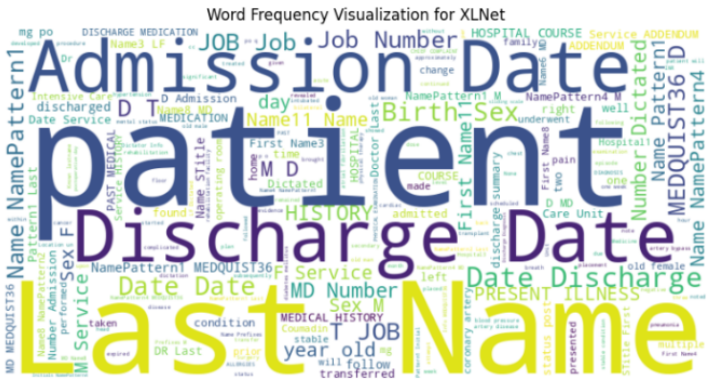

Figure 19: Word Frequency Chart

6.6 Discussion

The analyses of available summarization models with the help of ROUGE and BLEU scores clearly prove how efficient these solutions are in terms of capturing the necessary medical information. ROUGE-1, ROUGE -L and BLEU the metrics were demonstrated to have high results indicating that DistilBART is capable of generating brief and helpful resumes that hold significant information with high efficiency of 0.671, 0.608 and 0.600 correspondingly. A few assessments also relate to the quality and gist of the text where XLNet achieves a certain level of BLEU score of 0.634 with the ROUGE-2 of 0.519 proving that it is good at maintaining the coherent and detailed narration of the text. It is, therefore, evident from the findings that while both Pegasus and BERTSUM produce comparable performances, DistilBART and XLNet are ideal for summary generation in applications that require more precise information extraction.

7 Conclusion and Future Work

For Detailed Summaries: It is suggested to use distilBART because of high ROUGE scores and general and detailed summary length. For Concise Summaries: Overall, T5 is better since it has the highest BLEU score and yields summaries of moderate length as desired in the documents' preservation with concise and quality.As discussed above, that the data type we are operating

with currently is the text data, whereas the summarized data are missing. Currently, we are possessing 237 random data sets and to improve it more data sets will be included in the near future to enrich the model. Even though we have found the best model to date, the next task is to adapt it to clients' use. To this end we will use the model to deploy a web application using the FLASK framework to enable the recommendation of products. To improve the size of the application and apply isolation, Docker will be employed here. As for version control, we are going to use Git; in terms of testing and working with deployment, we will incorporate Jenkins. Lastly, the model can then be hosted on a cloud environment such as Amazon Web Services (AWS), Microsoft Azure or Google Cloud Platform.

References

[1] Ghadimi, A., & Beigy, H. (2023). SGCSumm: An extractive multi-document summarization method based on pre-trained language model, submodularity, and graph convolutional neural networks. *Expert Systems With Applications*, 215, 119308. https://doi.org/10.1016/j.eswa.2022.119308

[2] Zhang, M., Li, C., Wan, M., Zhang, X., & Zhao, Q. (2024). ROUGE-SEM: Better evaluation of summarization using ROUGE combined with semantics. *Expert Systems With Applications*, 237, 121364. https://doi.org/10.1016/j.eswa.2023.121364

[3] Searle, T., Ibrahim, Z., Teo, J., & Dobson, R. J. B. (2023). Discharge summary hospital course summarisation of inpatient electronic health record text with clinical concept guided deep pre-trained transformer models. *Journal of Biomedical Informatics*, 141, 104358. https://doi.org/10.1016/j.jbi.2023.104358

[4] Ozyegen, O., Kabe, D., & Cevik, M. (2022). Word-level text highlighting of medical texts for telehealth services. *Artificial Intelligence In Medicine*, 127, 102284. https://doi.org/10.1016/j.artmed.2022.102284

[5] Joshi, A., Fidalgo, E., Alegre, E., & Fernández-Robles, L. (2023). DeepSumm: Exploiting topic models and sequence to sequence networks for extractive text summarization. *Expert Systems With Applications*, 211, 118442. https://doi.org/10.1016/j.eswa.2022.118442

[6] Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E. P., Seehofnerova, A., Rohatgi, N., Hosamani, P., Collins, W., Ahuja, N., Langlotz, C., Hom, J., Gatidis, S., Pauly, J., & Chaudhari, A. (2023). Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. Stanford University. Available at: https://doi.org/10.21203/rs.3.rs-3483777/v1

[7] Keswani, G., Bisen, W., Padwad, H., Wankhedkar, Y., Pandey, S., & Soni, A. (2024). Abstractive Long Text Summarization using Large Language Models. *International Journal of Intelligent Systems and Applications in Engineering*, 12(12s), pp. 160–168

[8] Mohan, B., Archanaa, N., Kousihik, K., Kumar, R. P., Faheem, M., Daniel, V. J., & Kumar, J.D.T.S. (2024). Comparative Evaluation of Large Language Models for Abstractive Summarization. In *14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 59. DOI: 10.1109/CONFLUENCE60223.2024.10463521

[9] Landolsi, M.Y., Hlaoua, L., & BenRomdhane, L. (2023). Information extraction from electronic medical documents: state of the art and future research directions. *Knowledge and Information Systems*, 65, pp. 463–516. DOI: 10.1007/s10115-022-01779-1

[10] Bisen, W., Padwad, H., Wankhedkar, Y., Pandey, S., & Soni, A. (2024). Abstractive Long Text Summarization using Large Language Models. *International Journal of Intelligent Systems and Applications in Engineering*, 12(12s), pp. 160–168. Available at: www.ijisae.org

[11] Verma, P., Verma, A. and Pal, S. (2022) 'An approach for extractive text summarization using fuzzy evolutionary and clustering algorithms', *Applied Soft Computing*, 120, Article 108670. Available at: http://dx.doi.org/10.1016/j.asoc.2022.108670.

[12] Xie, Q., Bishop, J., Tiwari, P. and Ananiadou, S. (2022) 'Pre-trained language models with domain knowledge for biomedical extractive summarization', *Knowledge-Based Systems*, 252, Article 109460. Available at: http://dx.doi.org/10.1016/j.knosys.2022.109460.

[13] Silver, A.M., Goodman, L.A., Chadha, R., Higdon, J., Burton, M., Palabindala, V., Jonnalagadda, N., Thomas, A. and O'Donnell, C. (2022) 'Optimizing discharge summaries: A multispecialty, multicenter survey of primary care clinicians', *Journal of Patient Safety*, 18(1), pp. 58–63. Available at: https://doi.org/10.1097/PTS.0000000000000271.

[14] Torres-Parejo, Ú., Campana, J.R., Vila, M.A. and Delgado, M. (2021) 'On building and evaluating a medical records exploration interface using text mining techniques', *Entropy*, 23(10), pp. 1275. Available at: https://doi.org/10.3390/e23101275.

[15] Issam, K.A.R., Patel, S. and Subalalitha, C.N. (2021) 'Topic modeling based extractive text summarization', *CoRR*, abs/2106.15313. Available at: https://arxiv.org/abs/2106.15313.

[16] Mohan, G.B., Kumar, R.P., Parathasarathy, S., Aravind, S., Hanish, K.B. and Pavithria, G. (2023) 'Text Summarization for Big Data Analytics: A Comprehensive Review of GPT 2 and BERT Approaches', in R. Sharma, G. Jeon and Y. Zhang (eds.) *Data Analytics for Internet of Things Infrastructure*. Springer, Cham. DOI: 10.1007/978-3-031-33808-$3_1$4.

[17] Shi, J., Li, W., Yongchareon, S., Yang, Y. and Bai, Q. (2022) 'Graph-based joint pandemic concern and relation extraction on twitter', *Expert Systems With Applications*, 195(116), pp. 538. Available at: https://doi.org/10.1016/j.eswa.2022.116538.

[18] Touvron, H. et al. (2023) 'Llama 2: Open Foundation and Fine-Tuned Chat Models', *arXiv.org*. Available at: https://arxiv.org/abs/2307.09288.