BnClinical-Sum: Benchmarking Datasets for Bangla Long & Short Clinical Dialogue Summarization

by

Quazi Adibur Rahman Adib
21241056
Sanjana Binte Alam
20301455

A thesis report submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
January 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Quazi Adibur Rahman Adib
21241056

Sanjana Binte Alam
20301455

Approval

The thesis titled “**BnClinical-Sum**: Benchmarking Datasets for Bangla Long & Short Clinical Dialogue Summarization” submitted by

1. Quazi Adibur Rahman Adib(21241056)

2. Sanjana Binte Alam(20301455)

Of Fall, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 2024.

Examining Committee:

Supervisor:
(Member)

Dr. Farig Yousuf Sadeque
Assistant Professor
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

Ethics Statement

We accordingly state that the information provided in this thesis is based solely on our own research results. Every additional source implemented in this work has been sincerely recognized. Additionally, we attest that neither this thesis nor any of the parts of it has ever been turned in or presented in order to get a degree from any other university or educational institution.

Abstract

Despite significant improvements in the general-purpose text summarization task in the past decade, clinical conversion summarization is going through a tough time due to a lack of initiative to provide open-source datasets to the NLP community. In this work, we are presenting the first long and short Bangla Clinical Dialogue to Note Summarization datasets: **BnClinical-Sum**. Long conversations are detailed conversations with additional medical history. For the long dialogue dataset, we have accumulated around 207 pairs of full conversations and notes. Each note consists of in-depth discussions on previous medical histories, family medical records, and a wide variety of other topics. For the short dialogue version, our dataset consists of 1701 real-life short manually translated clinical conversations and their corresponding notes. The short dialogue dataset consists of subsets of long dialogue where each dialogue snippet addresses one sub-topic like previous medical histories, family medical records, etc. Those conversations are from 20 different categories like labs, assessments, plans, etc. Owing to demonstrating the efficacy of both datasets, we have trained our datasets on current state-of-the-art text summarization and text-to-text generative models to provide a solid benchmark for clinical conversion summarization tasks.

Keywords: ClinicalNLP; Dialouge2Note; mBART; Transformer; mLongT5; Benchmark; Summarization; Generation; Bangla; Dataset

Dedication

We would like to remember our wonderful parents, without whom we would be worthless, with all of our sacrifices and academic endeavors.

Acknowledgement

We convey gratitude to the Almighty for directing us to successfully complete our thesis without experiencing any significant obstacles. We also like to thank our peers, especially Niloy Farhan, Farah Binta Haque, Tahania Haseen, Jihadul Amin, Radiah Hassan, Farin Sharika Samia, Ishrat Tasnim, and Sami Anon for providing valuable suggestions and help in the overall thesis process. Finally, we would want to express appreciation for our parents' constant support, without which we might not have accomplished this success.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The automated process of extracting relevant medical information and generating comprehensive medical reports or summaries from spoken or written interactions between doctors and their patients is known as generating medical reports based on doctor-patient conversations using Natural Language Processing (NLP). This is one of the prominent use cases of text generation tasks. One of the simple block diagrams of the generative model is 10.2.

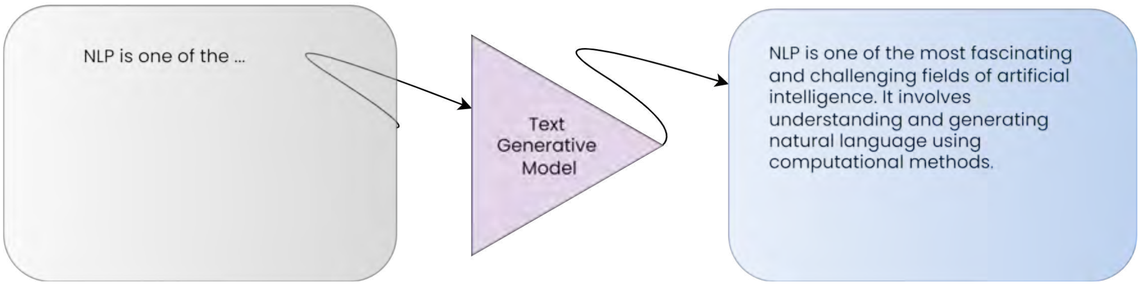

NLP is one of the ...
Text Generative Model
NLP is one of the most fascinating and challenging fields of artificial intelligence. It involves understanding and generating natural language using computational methods.

Figure 1.1: An inference diagram of Text Generative Model. Here, the left box contains an incomplete phrase as input. As an output, it completes that sentence and also adds other information in the left box. The triangle plays the role of a Text Generation Model which is responsible for generating relevant text in human language.

In order to improve healthcare documentation and simplify the healthcare process, the NLP-powered system analyzes the conversation, detects relevant clinical facts, organizes the information, and provides organized and coherent medical reports. The use of modern technology has completely changed how medical professionals do their work in the fast-changing medical environment. The use of Natural Language Processing (NLP) to generate medical reports from inquiries between doctors and patients is one such revolutionary development. A significant technology that enables the automatic extraction, analysis, and synthesis of medical information from the rich tapestry of discussions between medical professionals and their patients is NLP, an area of artificial intelligence. By providing timely insights and support to medical professionals throughout their interactions with patients, it helps in a number of unexpected ways. When making important decisions, it helps with real-time information retrieval by ensuring that clinicians have access to relevant medical information and patient history. This in-the-moment support may result in more precise diagnoses, individualized treatment strategies, and better patient outcomes. This ability to analyze data extends above specific patient interactions and enables medical facilities to make data-driven choices that can improve overall treatment, utilization of resources, and satisfaction among patients. This study explores the interesting intersection between NLP and medicine, demonstrating the latest innovations that are changing treatment by converting doctor-patient conversations into insightful, useful medical reports. We will look at how NLP is advancing medical research, enhancing patient experience, and transforming medical care records. We will also explore the ethical concerns that surround the use of this effective technology, ensuring that it is utilized properly and for the benefit of patients as well as doctors.

1.1 Motivation

Since our aim is to collect data and data benchmarking so we have collected our data from published scholarly articles which is originated from another country in English. Furthermore, we intend to create the dataset in Bangla, thus we will eventually convert the data manually. Due to lack of clinical reports in our country, we had to find ways to acquire our data from other countries. We attempted to do so for that reason. Additionally, we have made an effort to do this in order to prevent mistreatment in our country. Also, we sometimes get into trouble identifying our prescription or condition since we forget to keep track of our clinical reports. Thus, to prevent these issues and to possess a hope for a structured document in our country to facilitate the process.

1.2 Problem Statement

One of the rudimentary problems in Clinical NLP is the lack of a reliable dataset. In fact, in some cases, high-resource language is going through this problem. The central objective of this work is to develop datasets that consist of a pair of conversations between doctors and patients, and corresponding clinical notes of that conversation in Bangla. In this work, we are proposing two datasets for benchmarking this downstream task. As this domain is quiet, we do not have any previously reported dataset in the Bangla language. So, for the data structure, we have to rely on the English dataset.

Table 1.1: Currently Publicly Available English Dialogue to Note Dataset and Our Proposed Bangla Dataset

Publicly Available English Dataset			**Our Bangla Dataset**	
Dataset	**Size**	**Type**	**Dataset**	**Size**
MTS-dialogue	1701	Short	**BnClinical-Sum-Short**	1701
ACI-Bench	207	Long	**BnClinical-Sum-Long**	207
primock57	57	Long		

Here, we will utilize the reliable English dataset MTS-Dialogue and ACI-Bench [1, 2] and manually translate both of the datasets for the Bangla language. So that future researchers can find a good starting point for upcoming research.

1.3 Research Objective

The core objective of this work is to explore different settings of pipelines for different transformer-based models. This experimentation will aid us in understanding which design decision more aligns with the clinical domain.

- Explore different Transformer models.
- Explore different arrangements of the pipeline.
- Setup a Benchmark and Baseline performance for future research.
- Understand how text2text generative models works.
- Understand how the text summarization model works.

Chapter 2

Literature Review

The authors of this paper describe an elaborate process to create extensive sequence-to-sequence models to generate clinical notes from patient-doctor conversations. Treating the task as an abstractive summarization problem, they implement an encoder-decoder transformer model with a pointer-generator mechanism. The work, in particular, presents a number of modeling concepts, such as subword and multiword tokenization, prefixing target summaries with a chain of clinical facts, and training with a contrastive loss function. In order to effectively manage large input and output sequences, they also use flash attention during training and query chunked attention during inference. They used a substantial dataset for their tests, which show consistent accuracy. In addition, the authors also demonstrate how subword and multiword tokenization speed up model convergence while simultaneously improving accuracy. Faster training speeds are made possible by flash attention, which considerably increases batch size, while inference can be accomplished with every length of transcript according to query chunked attention. It is vital to remember that some parts, like the physical exam (PE) section, are unlikely to benefit from particular strategies for modeling since they are templated. [3]

Given the rising number of digital medical records in the healthcare industry, this paper [4] emphasizes the growing significance of automation in processing medical documentation. It comprehends how machine learning has been incorporated into medical decision-making systems, specifically in medical visualization, and the expanding significance of Natural Language Processing (NLP) methods in clinical situations. Because of their outstanding results on a variety of NLP tasks, including clinical NLP, large language models (LLMs), in particular transformer-based models, have drawn a lot of attention. Clinical applications have shown that there is potential for the use of prompt-based LLMs, especially for tasks like summarizing. The MEDIQA-Chat Tasks are also discussed in the study, with a concentration on summarizing and generating patient-doctor interactions. The findings indicate that certain models, like Clinical-T5-Sci, perform better than others at summarizing doctor-patient conversations and prediction of section headings.

This paper [5] indicated summarizing, categorizing, and creating patient-doctor discussions. They used a data-augmentation-first strategy (dialogue creation) taking the majority of their time. This strategy significantly enhanced model performance for all tasks. They produced impressive results using the adaptable BART architecture as the basis for their submissions, taking first place in Task C and placing strongly in other subtasks. They are the only team to provide stable and repeatable code for all three objectives, demonstrating their dedication to code stability and reproducibility. Their contributions and outstanding results highlight the value of augmented data and excellent code in clinical NLP tasks.

This article [6] describes a unique strategy known as the ”doctor-patient loop,” which makes use of Large Language Models (LLMs) in order to create conversation datasets of the highest possible standard. The results of their experiments show that they perform quite well while evaluated through an assortment of digital initiatives, including ROUGE, medical concept recall, BLEU, Self-BLEU, and others. Additionally, this study explores the possible application of cooperative LLMs in dataset development through the comparison of their preferred technique to Chat-GPT and GPT-4. This research extends its field by demonstrating innovative methods for collecting clinical conversation data and evaluating the standard and variety of created text.

This study [7] explores the differences and similarities between a number of transformer-based models, such as BioBART, Flan-T5, DialogLED, and OpenAI GPT-3, in the context of clinical dialog summarization, a field where such studies have not been extensively explored. The authors concentrate on summarizing both short as well as extended clinical conversations. They work with innovative methodologies and combined techniques to lessen hallucination of developed summaries. Each task is presented in three separate runs, and the authors use measures like ROUGE, BertScore, BLEURT, and multi-class accuracy to assess their performance. Additionally, they find that although the fact that reducing hallucination might generate better summaries, it could not always provide the best results because of metric biases that favor longer texts. This extensive research offers beneficial details about the efficiency of transformer-based models for clinical dialog summarizing.

The authors of this research focus on dialogue summarizing. They use two different pipelines: one that uses few-shot in-context learning (ICL) with the robust GPT-4 model and a conversation summarization model that has been fine-tuned. High scores in measures like ROUGE-1 F1, BERTScore F1, and BLEURT demonstrate the importance of these approaches in clinical note summarizing. Their techniques provide outstanding outcomes in these metrics. To provide a more thorough analysis, they furthermore employ expert annotations to look at the created summaries' accuracy significantly. The beneficial effects of using substantial language models and in-context learning for clinical conversation summarization are demonstrated by their findings, particularly when dealing with extensive clinical notes that reach beyond the limits of conventional models. [8]

This article presents the method that integrates conventional machine learning techniques, particularly Support Vector Machine (SVM), with the use of one-shot prompts implementing GPT-3.5 to generate representations that are different dialogues. Their model surpassed the usual standard results, demonstrating that it provided competitive performance. More specifically, their method surpasses the average performance of other task participants and gives an impressive overall score in dialogue summary. This work emphasizes the value of combining conventional machine learning with innovative language models in order to effectively categorize and summarize beneficial conversation. [9]

This article describes a method for the job of medical discussion summarizing. The authors enhance a LongT5 model on several different tasks at once, resulting in increased efficiency and less inaccuracies and hallucinations in the summaries that are created. They explore data augmentation using clinical named entity recognition tags but identify that it has an adverse impact on the quality of an overview. The study also analyzes various text creating techniques depending on note length. The outcomes indicate that the recommended method can improve the efficiency and impact of medical conversation summarization. Although the article presents beneficial information on summarizing medical conversations, it also relates with previous work on enhancing large language models for short tasks and evaluating data augmentation strategies for the accuracy of models.[10]

This article explores the field of clinical conversation summary, an essential application of NLP in the field of medicine. The objective of the study is to improve the proficiency of created medical chart notes by exploring combined techniques for summarization models. The study analyzes three alternative approaches, starting with a single summary model as a baseline and moving on to a collection of models that are specialized in various parts of the chart note. The last method involves layering or staging the results as they go through several summarization models. The results show that while ensemble models adjusted to certain areas generate better outcomes, the multi-layer method does not substantially improve efficiency. This research provides significant new perspectives on the opportunities of ensemble a summary models and opens up intriguing possibilities for clinical conversation summarization studies.[11]

The complicated job of medical dialogue summarizing is addressed in this study, which also introduces a distinctive system created for the Dialogue2Note Medical summarizing. The method, which uses a two-stage process for section-wise summary and focuses on choosing conversations that are semantically comparable and employing them as in-context examples for GPT-4, which has shown excellent outcomes. The article additionally looks at the effects of a variety of factors including the application of immediate structures and the overall number of in-context instances, on summarization performance. While emphasizing the benefits of few-shot prompting, the authors also note the obstacles in accomplishing the most effective summary length. The work shows extensive analysis of their performance, emphasizing both its advantages and disadvantages, and overall gives helpful insight into the creation of elaborate summarizing algorithms for medical conversations.[12]

Chapter 3

Task Definition

In this work, we are proposing a task for Bangla text generation. To solve this task, we have to create a Bangla text generative model where it can generate clinical notes from a dialogue:

3.1 Training

- Given: D, N;
  - Where $D := \{D_0....D_i\}$ set of Dialogues between patients and doctors
  - AND $N := \{N_0....N_i\}$ set of clinical Notes
- Here, we need to develop a text generative model model $F(D_j)$ which can produce a valid N_j which is not in $D_0...D_i, N_0...N_i$
- Characteristic of Valid N_k:
  - N_k must follow the syntax of a target language.
  - N_k must maintain clinical integrity.
  - N_k must follow the semantics of a target language.

3.2 Inference

- Input: D_i; D_i := Dialogue between patients and doctors.
- Output: N_i; N_i := Clinical Notes for D_i

Chapter 4

Challenges of Bengali Clinical Dialogue2Note Summarization

4.1 Challenges in Data Collection

Despite being one of the most widely spoken languages, Bangla has a very limited amount of workable text corpus for many downstream tasks. Clinical dialogue summarization is no exception to this. In fact, to our knowledge, we have not found any reported data for clinical dialogue summarization. The reasons behind this scarcity are:

Lack of Infrastructure. Most of the Bangladeshi hospitals, especially government hospitals, do not have digital data collection facilities. Moreover, medical reports and prescriptions are also handwritten. Also, hospitals do not have the facility to store conversions among patients and doctors. That is why it becomes sort of impractical for researchers to collect real-world data for Bangla language.

Privacy constraints. Ensuring Data Privacy is one of the most crucial things for clinical data collection and distribution. As we do not have any legal protocol for data sharing or selective distribution, hospital heists to share reports.

Lack of Motivation. This problem actually applies to every language. There are very limited amount of clinical dialogue datasets. In fact high resource language like English, have very limited accessible dataset (Yim et al., 2023, p.2)[2]. Owing to get a business advantage online scriber companies do not release datasets. Although, they have access to real-life data. This lack of motivation to create datasets for public use actually makes the data collection process difficult.

4.2 Challenges in Data Modeling

Currently available state-of-the-art generative models is not well suited for long text generation. On the other hand, a clinical report consists long texts. To process a long data we had to face lots of technical hardship. Also, for Bangla, we don’t have previously pre-trianed model in clinical documents. That is why it become difficult to train a model and get better result.

Chapter 5

Data Selection

As our core objective is to develop a dataset in Bangla from clinically approved sources and Benchmark those proposed datasets, we had to rely on the currently available English dataset which can be found in Table 5.1.

Table 5.1: Currently Available Clinically Approved English Datasets

English Dataset		
Dataset	**Size**	**Open**
MTS-dialogue[1]	1701	Yes
primock57[13]	57	Yes
aci-bench[2]	207	Yes
3M Health[14]	1342	No
Abridge[15]	6862	No
Augmedix[16]	500	No
emr.ai[17]	9875	No
Nuance[18]	802000	No

From this table, we can see that most of the datasets are not publicly available or open to use. There is a massive open data shortage in the clinical NLP regime. Table 5.1 shows that there are three publicly available datasets. From those datasets, we have selected MTS-dialogue[1] and aci-bench[2]. Both of the datasets are open to use. Also, Both of the datasets contain a handsome amount of data with respect to the clinical NLP domain.

5.1 Sources of BnClinical-Sum

BnClinical-Sum-Long. Among three publicly available datasets from table 5.1 there are two long datasets which are aci-bench and primock57. As aci-bench contains more data, that is why we have selected aci-bench for our **BnClinical-Sum-Long** dataset.

BnClinical-Sum-Short. Only publicly available short dialogue dataset is **MTS-Dialogue.** So we have no other options for short dataset.

5.2 Reason Behind Selecting Two Datasets

Automatically generating detailed notes can aid doctors in workplace. To develop such a system in Bangla. It is essential to have a certain dataset that can be helpful to NLP practitioners. That's why we have selected a long dialogue dataset.

Also, we can ignore the fact that till now Bangla text generation is at an early stage and it is still in the development phase. That is why sometimes generating long sequences can be really hard to generate. That is why to provide a sold starting point for the practitioners, we have also developed a short dialogue dataset.

Chapter 6

Bengali Short Dialogue2Note Summarization: BnClinical-Sum-Short

6.1 Data Sources

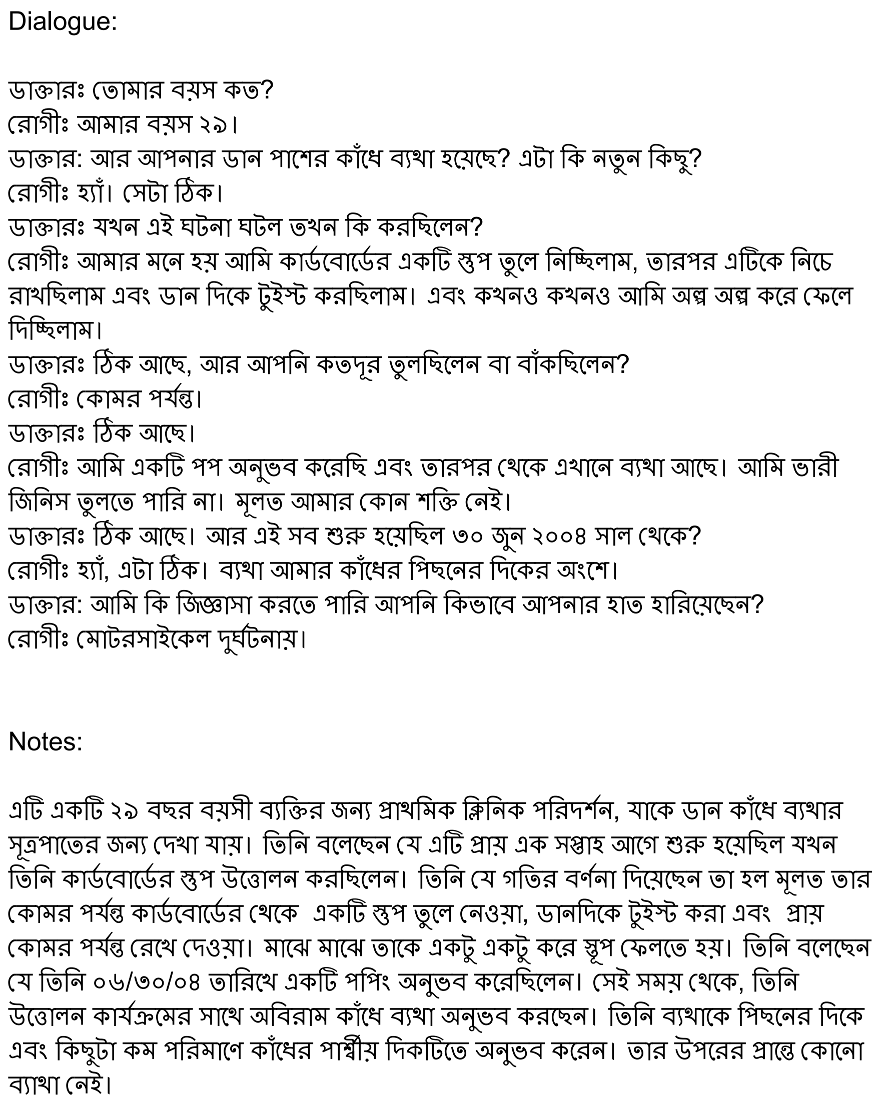
Our **BnClinical-Sum-Short** dataset is collected from the dataset of MTS-Dialogue (Abacha et al., 2023)[1]. This dataset contains dialogue-note snippets. For each short dialogue they have corresponding short notes which are clinically verified.

6.1.1 Data Creation

We've created our data from publicly accessible clinical notes. This data creation process comprises generating a medical report from doctor-patient conversations. We acquired the clinical notes along with summaries from the Mtsamples public collection, that gives established medical records. The six most frequently used note categories and specializations in the collection—General Medicine, SOAP (Subjective, Objective, Assessment, Plan), Neurology, Orthopedics, Dermatology, and Allergy/Immunology—are all covered by the chosen clinical notes. A total of 1,701 dialogue pairs and associated clinical note categories were included in the final MTS-DIALOG dataset. Data quality is corroborated from where we've collected the data.Conversations are rated by evaluators based on their agreement with annotation claims, topic applicability, and description.

6.1.2 Comparison with Real Data

The MTS-DIALOG dataset is constructed of authentic notes and simulated conversations that resemble doctor-patient interactions in order to prevent the confidential doctor-patient conversations. Through a blind review, we looked into how similar the MTS-DIALOG data was to actual discussions in order to examine the effects of depending too much on synthetic data. Differentiating between simulated and real data in the MTS-DIALOG dataset is ambitious. Although statistical study shows that MTS-DIALOG conversations have less speech flaws and pauses, medical specialists have noticed that the material feels real most of the time. Conversations that were clear, succinct, and comprehensible even with abrupt topic changes and conversational language were the reason why synthetic data was occasionally incorrectly identified as authentic. However, due to its probity, low speech flaws, and clarity, actual data was often mistaken for simulated data. This complexity focuses on the importance of the dataset as a starting point for model training and evaluation in real-world circumstances. In Bangla, each data point looks like in Figure 6.1



Dialogue:

ডাক্তারঃ তোমার বয়স কত?
রোগীঃ আমার বয়স ২৯।
ডাক্তার: আর আপনার ডান পাশের কাঁধে ব্যথা হয়েছে? এটা কি নতুন কিছু?
রোগীঃ হ্যাঁ। সেটা ঠিক।
ডাক্তারঃ যখন এই ঘটনা ঘটল তখন কি করছিলেন?
রোগীঃ আমার মনে হয় আমি কার্ডবোর্ডের একটি স্তুপ তুলে নিচ্ছিলাম, তারপর এটিকে নিচে রাখছিলাম এবং ডান দিকে টুইস্ট করছিলাম। এবং কখনও কখনও আমি অল্প অল্প করে ফেলে দিচ্ছিলাম।
ডাক্তারঃ ঠিক আছে, আর আপনি কতদূর তুলছিলেন বা বাঁকছিলেন?
রোগীঃ কোমর পর্যন্ত।
ডাক্তারঃ ঠিক আছে।
রোগীঃ আমি একটি পপ অনুভব করেছি এবং তারপর থেকে এখানে ব্যথা আছে। আমি ভারী জিনিস তুলতে পারি না। মূলত আমার কোন শক্তি নেই।
ডাক্তারঃ ঠিক আছে। আর এই সব শুরু হয়েছিল ৩০ জুন ২০০৪ সাল থেকে?
রোগীঃ হ্যাঁ, এটা ঠিক। ব্যথা আমার কাঁধের পিছনের দিকের অংশে।
ডাক্তার: আমি কি জিজ্ঞাসা করতে পারি আপনি কিভাবে আপনার হাত হারিয়েছেন?
রোগীঃ মোটরসাইকেল দুর্ঘটনায়।

Notes:

এটি একটি ২৯ বছর বয়সী ব্যক্তির জন্য প্রাথমিক ক্লিনিক পরিদর্শন, যাকে ডান কাঁধে ব্যথার সূত্রপাতের জন্য দেখা যায়। তিনি বলেছেন যে এটি প্রায় এক সপ্তাহ আগে শুরু হয়েছিল যখন তিনি কার্ডবোর্ডের স্তুপ উত্তোলন করছিলেন। তিনি যে গতির বর্ণনা দিয়েছেন তা হল মূলত তার কোমর পর্যন্ত কার্ডবোর্ডের থেকে একটি স্তুপ তুলে নেওয়া, ডানদিকে টুইস্ট করা এবং প্রায় কোমর পর্যন্ত রেখে দেওয়া। মাঝে মাঝে তাকে একটু একটু করে স্তূপ ফেলতে হয়। তিনি বলেছেন যে তিনি ০৬/৩০/০৪ তারিখে একটি পপিং অনুভব করেছিলেন। সেই সময় থেকে, তিনি উত্তোলন কার্যক্রমের সাথে অবিরাম কাঁধে ব্যথা অনুভব করছেন। তিনি ব্যথাকে পিছনের দিকে এবং কিছুটা কম পরিমাণে কাঁধের পার্শ্বীয় দিকটিতে অনুভব করেন। তার উপরের প্রান্তে কোনো ব্যাথা নেই।

Figure 6.1: Example of Short Dialogue and Notes in Bangla

6.2 Data statistics

The short dialogue dataset is divided into 20 categories. Each dialogue snippet deals with one category of problems. The distribution of categories can be found in Figure 6.2.

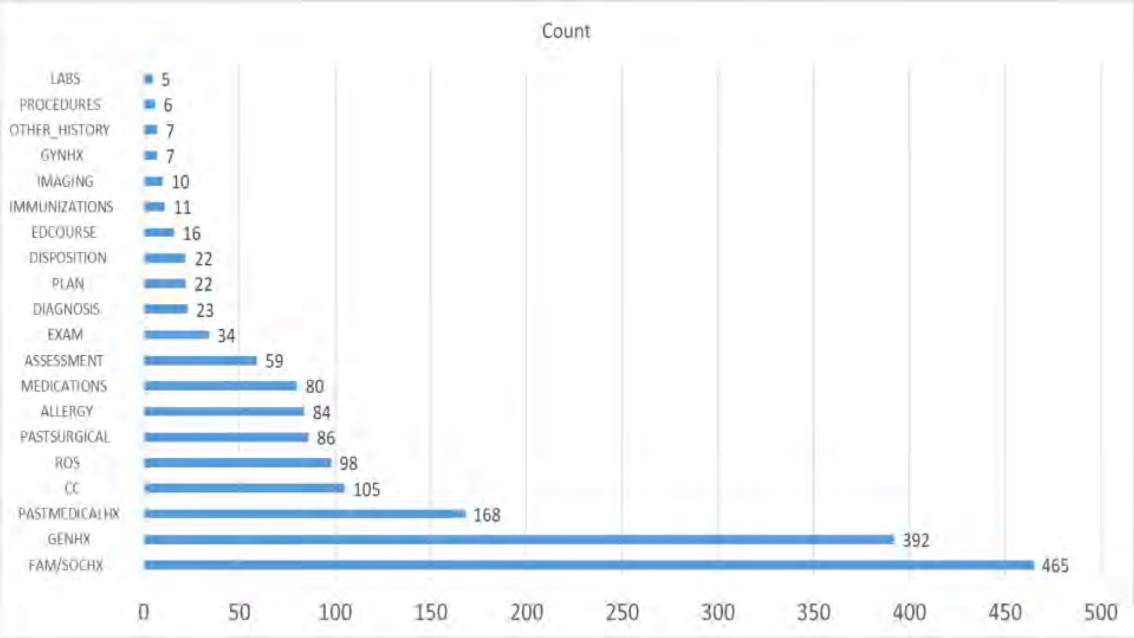


Figure 6.2: Class Distribution of Short Dialogue Dataset

From Figure 6.2, we can see that FAM/SOCHX has most of the data and LABS has minimal data.

Classwise note and dialogue token length can be understandable from Figure 6.3, 6.4 and Table 6.1.

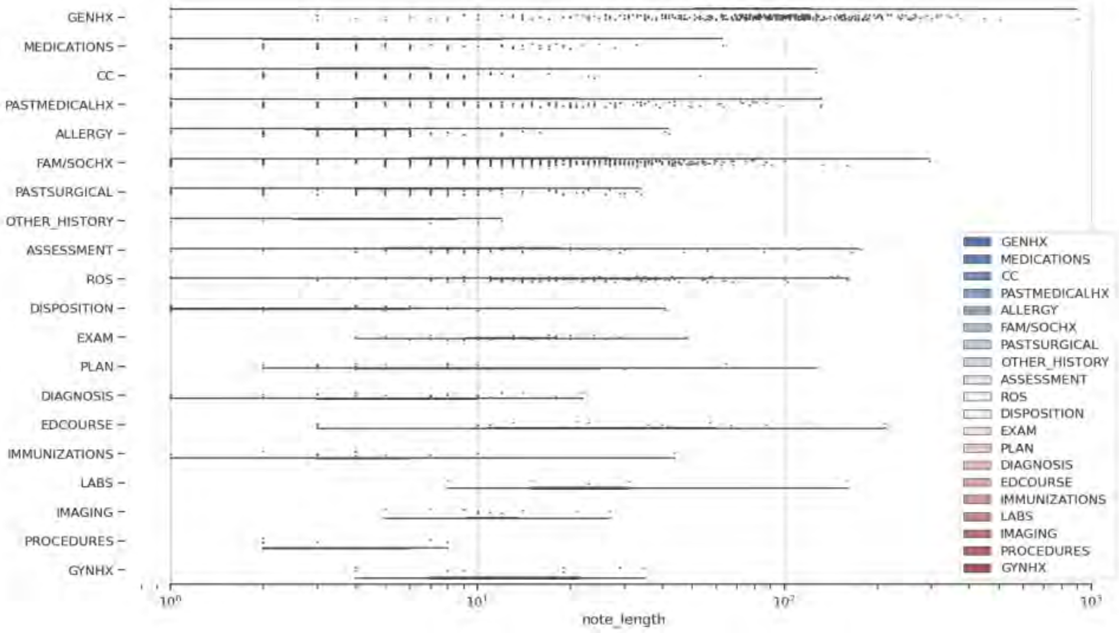


Figure 6.3: Classwise note token length

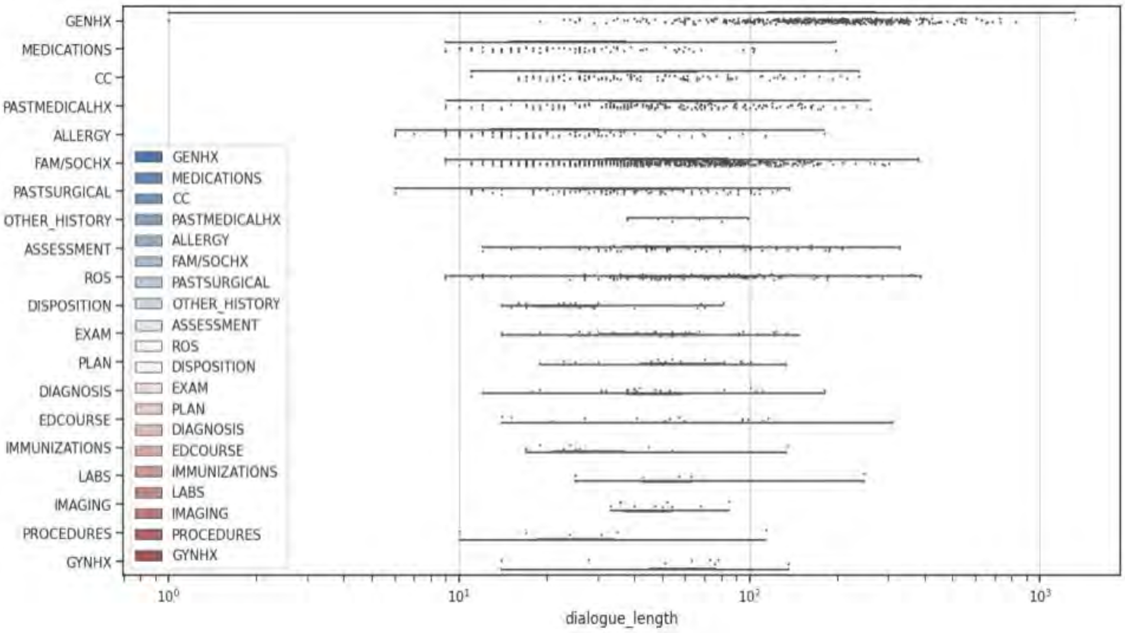


Figure 6.4: Classwise dialogue token length

Table 6.1: Note and Dialogue Token Size Summary (Short)

Index	Note Token Size	Dialogue Token Size
Count	1701	1701
Mean	96.34	94.087
std	148.75	102.7
Min	2	1
0.25%	15	82
0.5%	39	152
0.75%	111	307
Max	2521	3596

Chapter 7

Bengali Long Dialogue2Note Summarization: BnClinical-Sum-Long

7.1 Data Sources

Short Dialogue2Note Models are suitable for the initial survey. But for real-life use cases, it is essential to have a proper system that can handle the actual burden of a doctor. That is why we have decided to utilize one of the most comprehensive Dialogue2Note datasets which is created by Yim et al. (2023)[2]. We have manually translated this dataset for Bangla language to create our dataset.

7.1.1 Data Creation

In medical practice, clinical notes are indispensable documents produced by doctors, medical scribes, or individuals working together. The doctor can dictate these notes directly or with help, and they can be generated in an assortment of strategies that include formatting, precise information, and data to be included. With subsets like virtassist, virtscribe, and aci, the aci-bench corpus contributes an extensive picture of note-taking practices during doctor-patient interactions. In Virtassist, doctors talk to patients in a natural way and sometimes use specific terminology to incorporate virtual assistant features into consultations. Furthermore, transcripts made by medical professionals are included in virtscribe. These transcripts incorporate ASR and human transcripts, displaying a change of auditory strategies and interactions. ACI subset role-play prompts to simulate doctor-patient interactions. The automatically generated notes are then refined by the area of expertise. The complexity of creating clinical notes is highlighted by the collaborative efforts of medical professionals such as physicians, physician assistants, medical scribes, and clinical informaticists in creating these subgroups. It interprets how the consolidation of human knowledge, technical resources, and organized procedures results in accurate and thorough clinical documentation in a hospital encompassment.

7.1.2 Data cleaning and annotation

The final dataset included in this study came from simulated interactions for marketing demonstrations, where clinical notes were manipulated to include fictitious electronic health record (EHR) entries for the sake of authenticity. But often there was no clear connection between these entries and the actual discussion. The dataset did not link clinical notes to EHR inputs such as order codes, diagnostic codes, surveys, or vital signs. Annotation regulations were devised to determine dubious material in the notes compared to the actual discussion. These suggestions were designed to help note takers spot phrases that included facts that were not supported by the notes, such as treatment rationale unrelated to the discussion or content from fictitious EHR inputs. After text parts were identified, automated processing was used to eradicate them. In order to meet time limitations and a restricted first assessment, annotators found and fixed notes' mistakes to improve the quality of the dataset. Furthermore, discrepancies between the automatic speech recognition (ASR) transcripts and the clinical notes were resolved. Examples of ASR errors are identified for mismatches between transcript data and note content. Words such as "hydronephrosis" may be misspelled as "high flow nephrosis" in clinical notes, while names such as "castillo" may be misspelled as "kastio". It was the responsibility of the annotators to identify these discrepancies and provide correct information. Following annotation, correcting mistakes in notes and removing unsupported phrases from notes were part of the data processing phase. ASR transcripts were processed in two versions: original and ASR-corrected (manually edited) to assess the effect of ASR errors. After automated processing, encounters were manually reviewed to resolve remaining formatting errors and spelling errors. To provide better quality for further analysis and research, the dataset was enhanced by removing unjustifiable note content and resolving ASR issues.

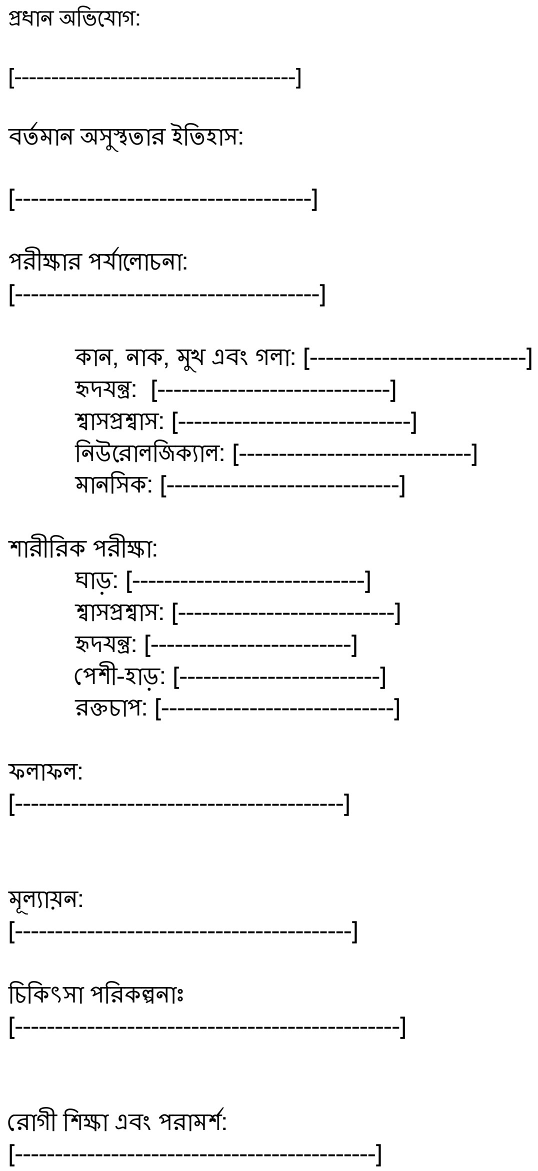
7.2 Structure of Notes

In this data, clinical notes are divided into several sections. Each section contains relevant clinical information about patients. In Bengali, The note looks like Figure 7.1. Details of units in English are:

CC. Here, CC means “Chef Complaint”. It indicates the exact characteristics of the patient's condition and its place of occurrence.

HPI. HPI refers to “History of Present Illness”. It includes information on the patient's symptoms, including their duration and intensity, as well as any relevant medical history.

PHYSICAL EXAM. It requires an extensive examination of a person's body to determine their overall health, detect any problems, and monitor how efficiently different parts of the body are functioning.



প্রধান অভিযোগ:

[-------------------------------------]

বর্তমান অসুস্থতার ইতিহাস:

[--]

পরীক্ষার পর্যালোচনা:
[---]

কান, নাক, মুখ এবং গলা: [------------------------------]
হৃদযন্ত্র: [--------------------------------]
শ্বাসপ্রশ্বাস: [--------------------------------]
নিউরোলজিক্যাল: [--------------------------------]
মানসিক: [--------------------------------]

শারীরিক পরীক্ষা:
ঘাড়: [-------------------------------]
শ্বাসপ্রশ্বাস: [------------------------------]
হৃদযন্ত্র: [----------------------------]
পেশী-হাড়: [----------------------------]
রক্তচাপ: [--------------------------------]

ফলাফল:
[---]

মূল্যায়ন:
[--]

চিকিৎসা পরিকল্পনাঃ
[--]

রোগী শিক্ষা এবং পরামর্শ:
[---]

Figure 7.1: Example of Long Notes in Bangla

RESULTS. The results acquired from diagnostic procedures or imaging scans, including biopsies, MRIs, and X-rays, besides other medical testing. These findings frequently show if particular illnesses, disorders, or abnormalities are present or not.

ASSESSMENT. Exams, tests, observations, and conversations are all part of evaluating a patient's health in order to make a diagnosis, plan a course of therapy, or monitor a patient's condition.

PLAN. The term ”plan” describes the suggested strategy of actions that a medical professional plans to conduct after evaluating the patient’s condition.

HISTORY. ”History” usually refers to the medical history or medical file of the patient. It includes a thorough description of all of a patient’s previous and present medical diseases, sicknesses, surgeries, prescription medication, allergic reactions, family members medical history, their way of life, and other important health-related information.

PE. A physical examination (PE) is a comprehensive assessment carried out by a healthcare provider to carefully examine a patient’s health. It means giving attention to, sensing, touching, and hearing various physical areas. The intent of this examination is to recognize health issues, provide appropriate medical recommendations, monitor general health, and diagnose diseases. It involves monitoring vital signs, assessing the health of the body, and carrying out particular procedures in response to symptoms or past medical records.

FINDINGS. Information on a patient’s health state, symptoms, diagnosis, and existence of any abnormalities that is obtained by lab results, physical examinations, imaging scans, or medical tests.

ASSESSMENT AND PLAN.A crucial element of medical documents such as patient records, the ”Assessment and Plan” summarizes the findings of a healthcare provider’s evaluation of a patient’s health state and suggested method of treatment. The Plan contains indicated operations, such as treatments, substances, or follow-up calls, while the Assessment includes a summary of the medical condition based on the data collected. This established system offers a consistent strategy for focusing on the patient’s health issues, simplifies patient care, and improves clinicians communication with one another.

ORDERS. ”Orders” are guidelines for a patient’s medication, substances, tests, or procedures that are given by a doctor. Staff nurses or other healthcare professionals involved in the patient’s care generally execute these orders, which are recorded in the patient’s medical records or electronic health system.

7.3 Data statistics

From Table 7.1, we can see that the count is quite low but the token size for both notes and dialogue is high.

Table 7.1: Note and Dialogue Token Size Summary (Long)

Index	Note Token Size	Dialogue Token Size
Count	207	207
Mean	1115.76	2649.24
std	445.82	896.87
Min	362	442
0.25%	902	2020.5
0.5%	1057	2487
0.75%	1278.5	3076
Max	4497	6142

From Table 7.1, we can say that the overall corpus is larger in terms of token size. Also, the 25th percentile is only 902. So maximum data points are contains a lots of tokens.

Chapter 8

Drawbacks of Google Translator for Clinical Data Translation

Initially, we used Google Translator to refactor our data from English to Bangla. In our observation, we noticed certain grammatical and semantical issues. For example,

- For certain phrases, it provides their literal meaning.
- Sentence structures are not always maintained.
- It incomprehensibly uses the word orientation.
- It also adds irrelevant sentences.
- It adds unnecessary phrases that are not aligned with Bangla languages.

8.1 Quantitative Analysis: Translation Quality

Owing to providing a quantitative basis for this work, we have compared Google-translated texts respected to our manually translated text. Our results show that, in ROUGE, the performance was not good enough. In case of BLEU, it was quite good. But as we are dealing with a sensitive domain like clinical NLP where people's life involved. We should focus on precious work.

Metric	**Performance**
BLEU	0.73
ROUGE	0.26
BLEURT	0.6

That is why, we have manually modified those issues to follow our annotation guidelines. Our detailed annotation guideline is presented in the next chapter of this report.

Chapter 9

Data Annotation Guideline

To our knowledge, there is no data available for Bangla clinical dialogue2note tasks. Also, a proper annotation guideline for data translation is also unavailable. Owing to making this work standardized, we propose an annotation guideline for clinical document translation from one language to another. We tried to make this thing as comprehensive as possible. So, this guideline can be used for every language.

We judge a clinical dialogue and notes as suitable content for the clinical environment based on the accurate linguistic features and adaptation to the clinical environment of a certain country. In our case, we are working with Bangladeshi culture. To measure dialogues and notes, we rely on the grammatical correctness of a given text and maintain all clinical integrity. The core objective is to focus on:

- Dialogue: Does the Dialogue resemble the spoken culture of the target language?
- Notes: Does the Notes resemble the written documentation style of the target language?
- Does both Dialogue and Notes maintain clinical integrity?

9.1 Pipeline for Data Translation

Initially, we translated our text using Google Translator. After observing the translation, we have found some major discrepancies. To make the data suitable for Bangla clinical dialogue to note generation tasks, we have focused on six points to ensure uniformity and reproducibility, we have set up certain rules for Data Translation.

Changing Numbers. The data we obtained from Google Translator had an issue. Certain numbers were in English. We need to change them because we are working with Bangla datasets. Thus, all of the numerals have been converted to Bangla. Like Table 9.1.

Table 9.1: Fixing Changing Numbers

Before	**After**
যিনি একজন 50 বছর বয়সী পুরুষ।	যিনি একজন ৫০ বছর বয়সী পুরুষ।

Word Orientation. Another problem we encountered is that some words are reversed in Google translator which does not make any sense. We worked on those words and constructed the dataset by forming meaningful sentences with the correct words. Like Table 9.2.

Table 9.2: Fixing Word Orientation

Before	**After**
রোগীর রিপোর্ট তার বাবা-মা উভয়েরই উচ্চ রক্তচাপ ছিল।	রোগী জানান তার বাবা-মা উভয়েরই উচ্চ রক্তচাপ ছিল।

Sentence Structure. There were some sentences in the data that we got from the translator that were not translated correctly, resulting in no meaningful sentences. We endeavored with those sentences and created the dataset by arranging the sentences according to the structure of the sentences. Like Table 9.3

Table 9.3: Fixing Sentence Structure

Before	**After**
আপনি কেমন আছেন যা আপনাকে নিয়ে আসে	আপনি কেমন আছেন ? কেন এসেছেন আজ ?

Abolishing Irrelevant Sentences. Some irrelevant terms that are either not needed to create a medical report or from which no information can be obtained have been left out of the dialogue. The sentences that are stated only to give the conversation additional complexity have been left out. Like Table 9.4

Table 9.4: Abolishing Irrelevant Sentences

Before	**After**
ডাক্তারঃ ওহ, আমি দুঃখিত। কি দারুন !	ডাক্তারঃ আমি দুঃখিত।

Unnecessary Phrases. There are frequently some redundant phrases used while we converse. Likewise, the translator translated several unnecessary phrases from the doctor-patient conversation. Thus, in order to generate our dataset, we excluded those extraneous phrases. Like table 9.5

Table 9.5: Abolishing Unnecessary Phrases

Before	**After**
ডাক্তারঃ হাই, আপনি কেমন আছেন !	ডাক্তারঃ আপনি কেমন আছেন !

9.1.1 Maintaining clarity and clinical terminology

To ensure clinical integrity, we have not changed, altered, or paraphrased any medical and clinical term. Moreover, we have not improvised any clinical decision.

Chapter 10

Benchmarking Model Descriptions: Architectures and Pre-training Protocols

10.1 Overview of Benchmarking Models

Table 10.1: Brief Description of Models for Benchmarking

Model	Developed By	Dataset	Training Strategy
mLongT5	Google Research	mC4	Principle Sentences Generation + Local Attention + Transient Global Attention + UL2 + LongT5
CrossSum	BUET CSE	CrossSum	Vanilla T5 + Multistage Language Sampling
CrossSum__Enhanced	BUET CSE	CrossSum	Vanilla T5 + Multistage Language Sampling
mBART Large	FAIR	ML50	Vanilla T5 + Noising Function to Maximize Loss Function
mT5 Multilingual XLSum	BUET CSE	XL-Sum	Vanilla T5
BanglaT5	BUET CSE	Bangla2B+	Vanilla T5

For our experimentation, we have used 6 different models. Here, all of the models are inspired by Transformer models[19]. In upcoming sections, we will discuss basic of transformers, and strategical difference among our selected models from Table 10.1.

10.2 Basic Overview of Transformers

In 2017, Google researchers introduced a groundbreaking deep learning architecture known as the Transformer [19], in their paper titled "Attention Is All You Need." This neural network model boasts remarkable versatility and can be applied to a wide array of natural language processing tasks, including but not limited to language translation, text summarization, and language comprehension. The Transformer model's foundation lies in the concept of self-attention, which empowers it to assign varying levels of importance to different segments of the input data when making predictions. This model harnesses multiple attention heads to understand diverse relationships among input tokens, enabling it to capture intricate patterns within the data.

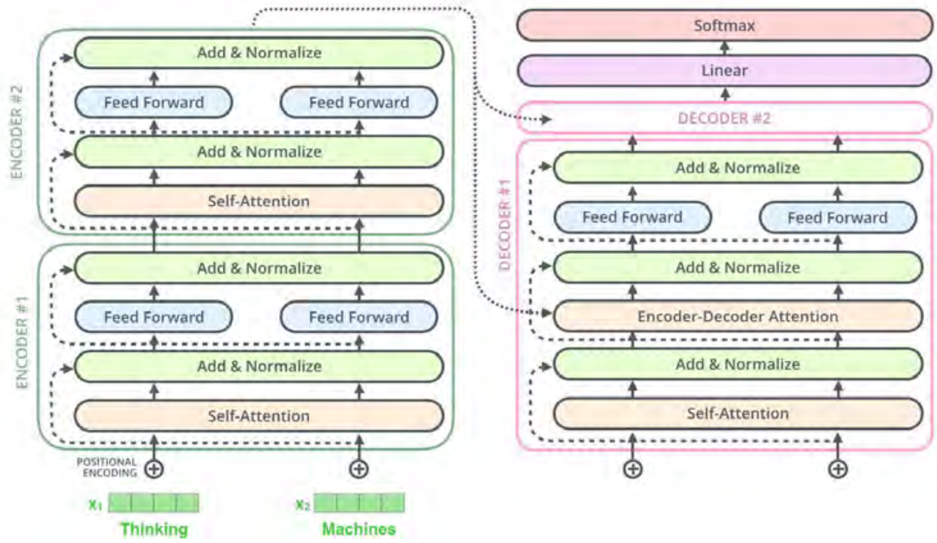


Figure 10.1: Basic Overview of Transformer. [20]

10.2.1 Architecture

Transformer architecture consists of two different types of network blocks which are the encoding blocks and decoding blocks. As for our analysis, we have used the base version of each model. So, our model consists of 12 encoder blocks and decoder blocks, with about 220 million parameters.

Encoder-Block

The encoder operates in an important part in the process of creating medical reports from discussions between doctors and patients. It begins by processing unstructured input data, which is frequently displayed as text or audio conversations, and then goes on to extract important factors, such as medical entities and conversational context. The encoder then transforms all of this data into numerical embeddings that contain semantic and contextual information, assuring that the medical report effectively represents each aspect of the discussion. Its results create an organized framework on which specified medical reports can be developed, responding to medical experts to offer reliable and proficient medical treatment. These reports require important sections like patient information, medical records, diagnosis, along with treatment recommendations.

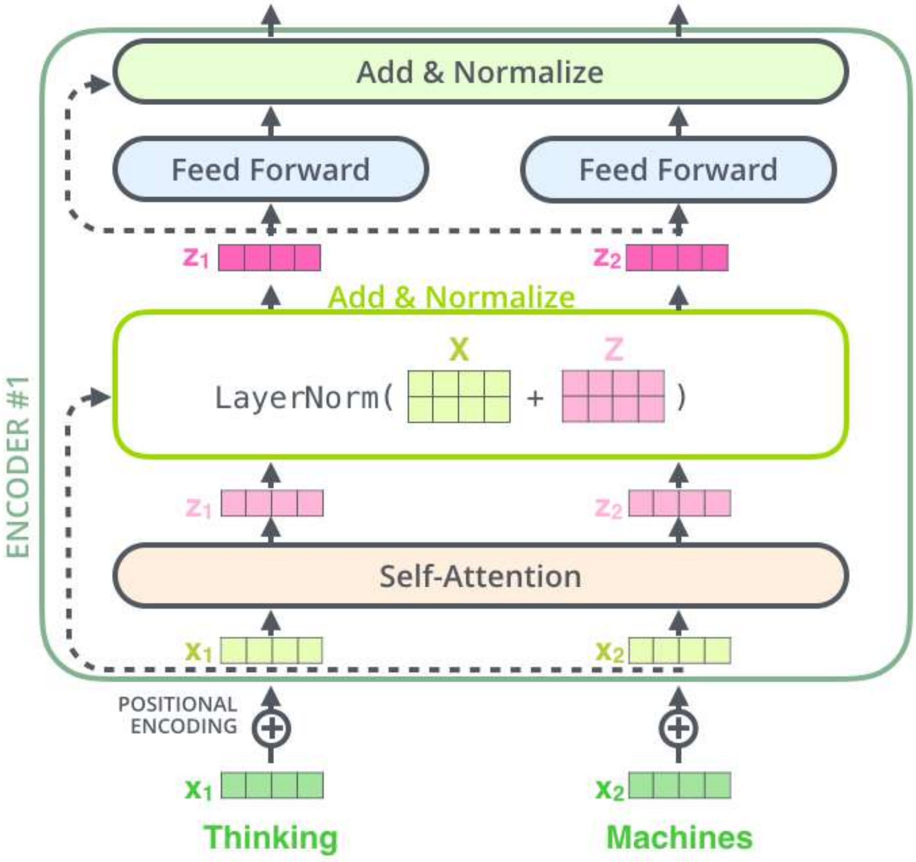


Figure 10.2: Architecture of Encoder. [20]

Decoder-Block

The decoder, which converts the structured data acquired by the encoder into accessible and medically relevant reports, is an essential part of the process of producing medical reports from doctor-patient interactions. In addition to producing text, the decoder puts together data into established report segments, adjusts results based on different medical specialities, and provides concise clarifications in simple language. Furthermore, to be completely accurate and clear, the generated reports must be understandable to both patients and doctors. This intricate performance subsequently enhances medical treatment and enables more proficient doctor-patient communication.

Self-Attention Mechanism

The self-attention mechanism is an essential part of Natural Language Processing (NLP) models when it involves creating medical reports to provide information between a doctor and a patient. It gives the machine the ability of focusing on important conversational threads while recording complicated word-and-phrase connections. This approach enables the model to analyze the proportional significance for multiple dialogue components, presenting the extraction of relevant clinical information including signs and symptoms, diagnosis, and medication options. The self-attention mechanism optimizes precision and understanding of context of the established medical reports by dynamically modifying its attention during report creation, to assure that they offer medical experts insightful and clinically appropriate knowledge.

Positional Encoding

Positional encoding is an important part of Natural Language Processing (NLP) models which generate medical reports from interactions among doctors and patients. It explores the challenges of keeping word order information in text data sequences. The sequencing of information in medical interactions is usually vital to precise diagnosis as well as treatment planning. Each sentence or object in the dialogue is given a distinct value in numbers through positional encoding, implying its precise position in the sequence. This enables NLP models to be able to take consideration for both the content and the word order, assuring that the created reports on health correctly portray the dialogue's sequential order. The records' context-awareness and integration are additionally enhanced by positional encoding, which further improves their medical significance as well as utility for clinicians.

10.3 mBART-Large-50

mBART is developed by the Facebook AI Research group. The main architecture of this model is based on transformer architecture. It is basically a denoising model like Transformers. The main difference between Google's transformer and Facebook's BART actually lies in the training process and multilingual capability. From the architectural point of view. mBART contains an additional layer-normalization layer on top of both the encoder and decoder, which enhanced training stabilized at FP16 precision.

10.3.1 Training Process of mBART-Large-50

The core fundamental of mBART-Large-50[21] is primarily inspired from the seq2seq modeling scheme where it uses a denoising autoencoder. It's central objective is to generate original text X from a noisy text $g(X)$ where $g(\cdot)$ is a predefined nosing function. To calculate and maximize loss using Eq: 10.1.

$$L_\theta = \sum_{D_i \in D} \sum_{X \in D_i} \log \left(X | g(X); \theta \right. \tag{10.1}$$

Noising Function. Noising function $g(\cdot)$ basically injects noise in two ways. Firstly it randomly masks like [19] and they introduce change in order in the text. For random masking, they have masked around 35% of the words[22].

Dataset. For pertaining their model they have used ML50 Benchmark dataset. It consists of 50 different languages. It consists of high-resource languages like French to extremely low-resource languages like Gujarati. Also, they have utilized WMT, IWSLT, TED58, OPUS, WAT, LauraMartinus, ITB, and FLORERS datasets [21].

10.3.2 Multilingual Translation Model Variants

There are several variations of the multilingual model for mBART-large-50. From those, we have many-to-many generative settings. Also, they have proposed many-to-one, one-to-many, and many-to-many.

10.3.3 Multilingual Finetuning

For multilingual fine-tuning, they have collected pairs of different languages in bi-texts format. They augmented both instances of a given pair and developed a dataset. After that they finetuned a pretrained mBART to enhance multilingual capabilities[21].

10.4 CrossSum and CrossSum-Enhanced

CrossSum and CrossSum-Enhanced both are mainly pre-trained versions of the classical transformer model[23] with Multistage Language Sampling. To train their model, they have used the CrossSum Benchmark dataset which consists of 1.68 million article-summary samples in 1,500+ language pairs.[23]

10.4.1 Training Process

For data modeling and training they followed the strategy of mT5 model[24]. Apart from this they have also introduced **MLS** approach. As their proposed CrossSum dataset is highly imbalanced.

Multistage Language Sampling (MLS)

MLS is fundamentally a probability smoothing technique for upsampling in multilingual pertaining settings. The basic idea is to fit repeated samples in each batch from low-resource language to make each batch somehow stratified.

$$p_i = \frac{\sum_{j=1}^{n} c_{ij}}{\sum_{j=1}^{n}\sum_{k=1}^{n} c_{ij}}; \forall i \in \{1, 2, ..., n\} \tag{10.2}$$

$$q_i = \frac{p_i^{\alpha}}{\sum_{j=1}^{n} p_j^{\alpha}}; \forall i \in \{1, 2, ..., n\} \tag{10.3}$$

$$p_{j|i} = \frac{c_{ij}}{\sum_{k=1}^{n} c_{ik}}; \forall j \in \{1, 2, ..., n\} \tag{10.4}$$

$$q_{j|i} = \frac{p_{j|i}^{\beta}}{\sum_{k=1}^{n} p_{k|i}^{\beta}}; \forall j \in \{1, 2, ..., n\} \tag{10.5}$$

The overall sampling process follows those four probability equations 10.2, 10.3, 10.4, and 10.5. Here, c_{ij} refers to the data count from language i to language j translation. and α, β are smoothing factors.

During batching, they leverage q_i and $q_{j|i}$ to sample data from each batch to overcome data imbalance. After that, they train the model like mT5[24].

Dataset One of the main contributions of [23] is the dataset which consists of 1.68 million article-summary samples in 1,500+ language pairs. To make data coverage larger they have used 'induced pairs' and 'implicit leakage.'

10.5 mLongT5-base

As long dataset has input and output, the token count is quite large. That is why It was essential for us to incorporate a model that can inherently manage long inputs and outputs. Although it is suitable for long text, surprisingly the model architecture is the same as mLongT5. The main difference lies in its attention mechanism, unlike vanilla transformers. It employed both local and global attention to take. Also, they have used Unifying Language Learning Paradigms which is based on Mixture of Denoising [25]. To understand how mLongT5 works, it is essential to know about how LongT5 works. In the upcoming subsection, we will discuss it.

10.5.1 LongT5 Basics

Architecture. LongT5 mostly used T5[24] as a foundation. The only difference lies in the encoder block where they have used certain changes in attention block to handle long sequences. They have introduced two changes first one is Local Attention and the other change is in the Transient Global Attention (TGlobal) block.

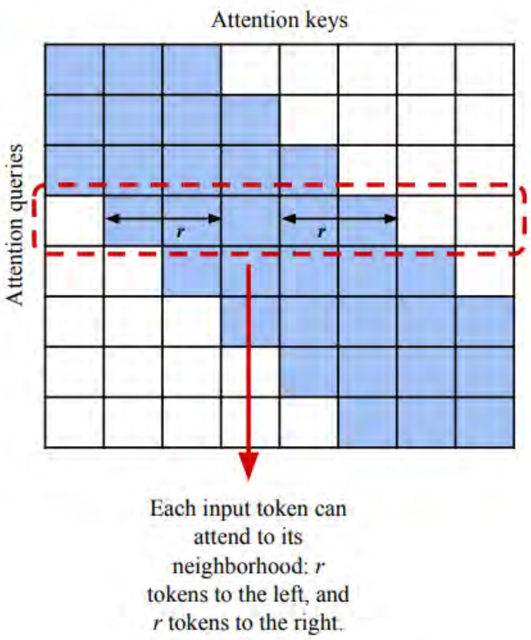

Attention keys
Attention queries
r
r
Each input token can attend to its neighborhood: r tokens to the left, and r tokens to the right.

Figure 10.3: Process diagram of Local attention in LongT5 Architecture [26]

Local Attention. It replaced the classical vanilla T5's encoder with a sparse sliding window for local attention. That actually selects tokens from a given radius (r). In their experimentation, they have seen that the ideal value for the radius is 127. The time complexity of the overall process is $O(l \times r)$. Here, l refers to the length of each input string. The basic idea can be understood from Figure 10.3.

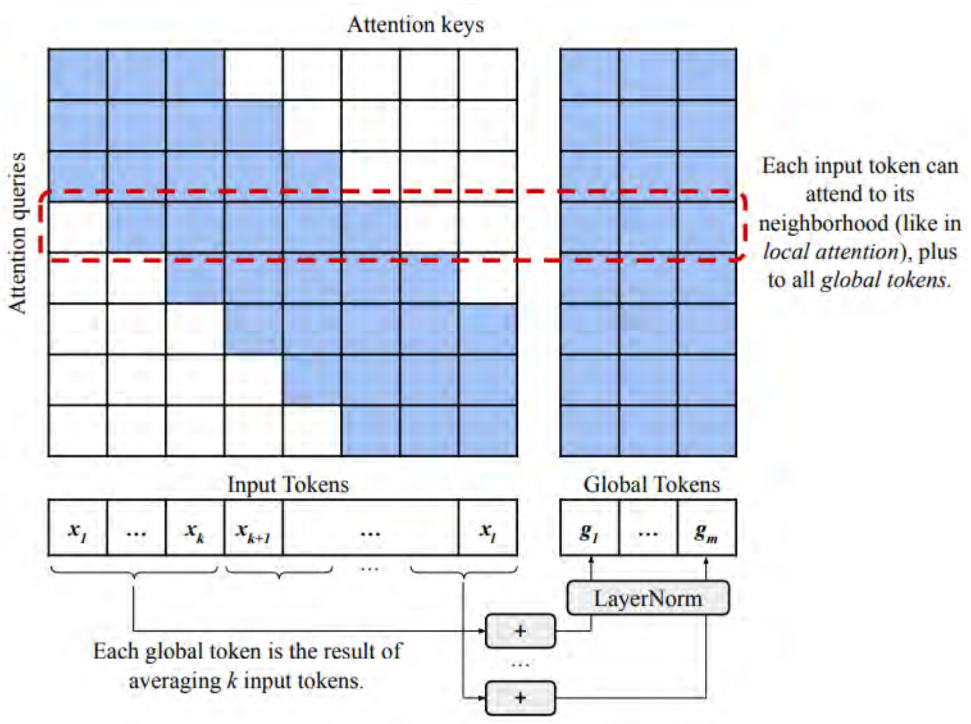

Attention keys
Attention queries
Each input token can attend to its neighborhood (like in *local attention*), plus to all *global tokens*.
Input Tokens
x_1
...
x_k
x_{k+1}
...
x_l
Global Tokens
g_1
...
g_m
LayerNorm
+
...
+
Each global token is the result of averaging *k* input tokens.

Figure 10.4: Process diagram of Transient Global Attention in LongT5 Architecture [26]

Transient Global Attention (TGlobal). Another improvement in the LongT5 model is the introduction of Transient Global Attention (TGlobal) mechanism. The core idea of TGlobal is to establish connections between local tokens with global tokens. Here, they have computed attention with both a fixed k-sized block in the input token and global tokens. This encapsulates both important information of both local and global information. As knowing a handsome amount of information is necessary for generating a long sequence. The overall idea is understandable from Figure 10.4. In this process, the input sequences are divided into k fixed blocks. So, to calculate model needs to iterate $\frac{l}{r}$ time. For calculating only TGlobal, it requires $O(l \times \frac{l}{k})$ time where l is sequence length and k is the block size. Ideally, the block size is $k = 16$.

Overall Time Complexity and New Parameter Count. To execute both the mechanism the overall complexity becomes $O(l \times r + l \times \frac{l}{k}) = O(l \cdot (r + \frac{l}{k}))$. For parameter changes, Firstly, T5-style relative position biases indicate the separation between the block of an input token and the block of every global token it is processing. Another parameter edition is T5-style layer normalization parameters that normalize the embedding of every global token [26].

Unifying Language Learning (UL2). Apart from local attention and Transient Global Attention, another optimization LongT5 introduced is leveraging the idea of UL2 [25].

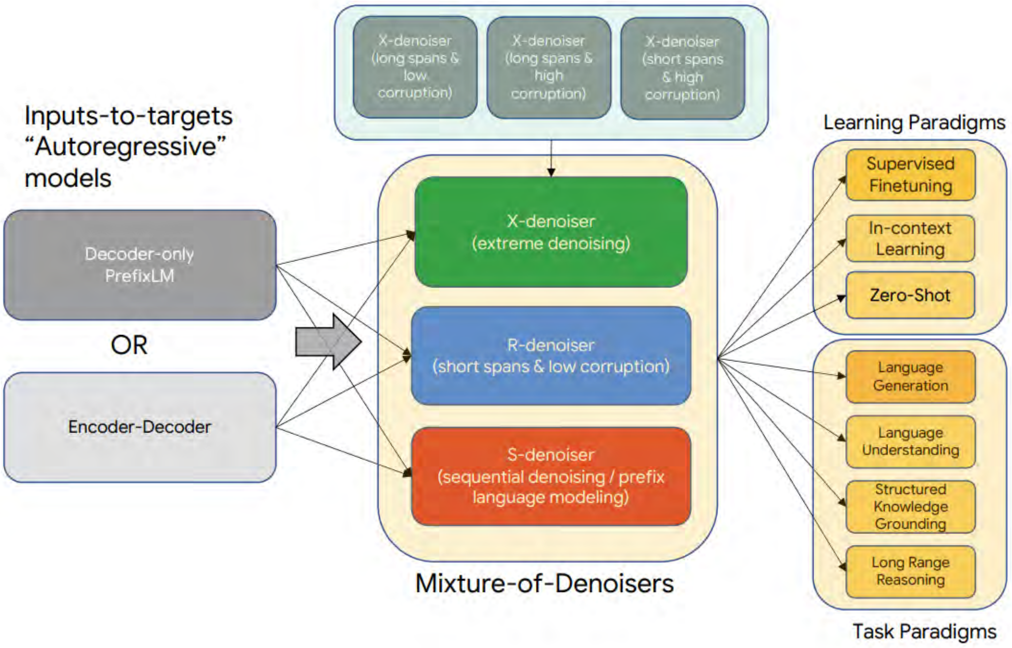

X-denoiser (long spans & low corruption)
X-denoiser (long spans & high corruption)
X-denoiser (short spans & high corruption)
Inputs-to-targets "Autoregressive" models
Decoder-only PrefixLM
OR
Encoder-Decoder
X-denoiser (extreme denoising)
R-denoiser (short spans & low corruption)
S-denoiser (sequential denoising / prefix language modeling)
Mixture-of-Denoisers
Learning Paradigms
Supervised Finetuning
In-context Learning
Zero-Shot
Language Generation
Language Understanding
Structured Knowledge Grounding
Long Range Reasoning
Task Paradigms

Figure 10.5: Overview of Unifying Language Learning (UL2). [25]

The overall workflow of UL2 is described in Figure 10.5. From the figure, we can that each autoregressive model learns from three different denoising schemes. Combining those three schemes is called Mixture-of-Denisers.

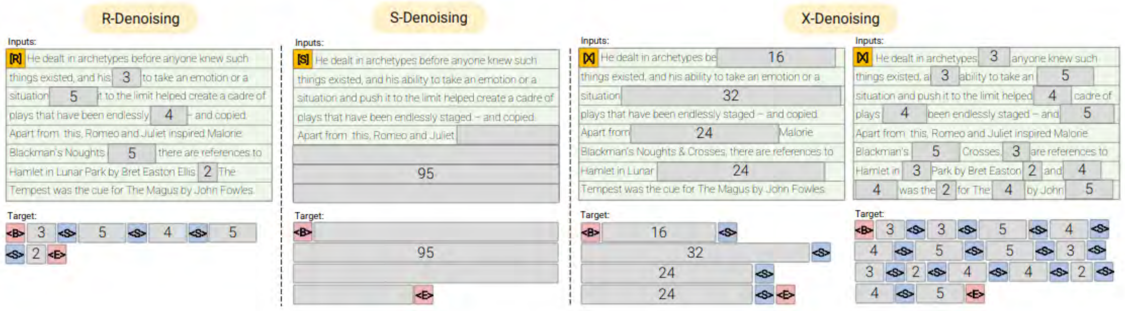


Figure 10.6: Denoising Scheme of Unifying Language Learning (UL2). [25]

UL2 introduces three different types of noise schemes. Firstly, X, R, and S. Here, the X denoising module actually helps the model to learn to generate long sequences. R-denoiser is actually the same as the classical method [19]. Finally, the S-denoiser helps the model to generate sequential output from a prefix like Figure 10.6.

10.6 Multilingual-XLSum

The backbone network for Multilingual-XLSum is also mT5 or Transformer[24]. This model is an mT5 model which utilizes the model architecture and it is finetuned with [27] dataset. This dataset has access to 44 different languages. The dataset consists of a total 1005292 samples. The main source of the dataset is BBC News[1]. For summary, they also rely on BBC's summary. As BBC also provided a summary in one/two sentences at the beginning of each article.

10.7 BanglaT5

BanglaT5[28] model utilizes the transformer model[19]. The training strategy is similar to the classical English model. For development of BanglaT5 they have uses Bangla2B+ dataset[29]. According to their report, this model is suitable for Machine Translation, Text Summarization, Question Answering, Multi-turn Dialogue, News Headline Generation, and Cross-lingual Summarization tasks. Like Multilingual-XLSum pertrained model BanglaT5 actually developed using Bangla2B+ dataset on vanilla mT5 model [24]

[1] https://www.bbc.co.uk/ws/languages

Chapter 11

Benchmark

For both of the datasets, we have followed sort of same type of data preprocessing and fine-tuning strategy. We have mainly, focused on working with two types of off-the-shelf models. The first one is generic `Text2Text Generation Models` and the second one of `Summarization Models`.

11.1 Dataset Split

For this experiment, we have divided our data into three sets. Out train:validation:test set ratio is 64:16:20 for both of the tasks.

11.2 Fine-tuning Protocol

Short Dataset Fine-tuning. We have fine-tuned our model with 5 epochs and the batch size was 2 as each data contains around 1024 input tokens and 512 output tokens.

Long Dataset Fine-tuning. We have fine-tuned our model with 10 epochs and the batch size was 1 as each data contains around 4096 input tokens and 900 output tokens. We had to keep our batch size small due to hardware limitations.

11.3 Pre-processing Pipeline

Preprocessing is one of the most fundamental steps for any natural language processing task. In this work, we have employed a generic common pipeline for preprocessing which consists of three steps:

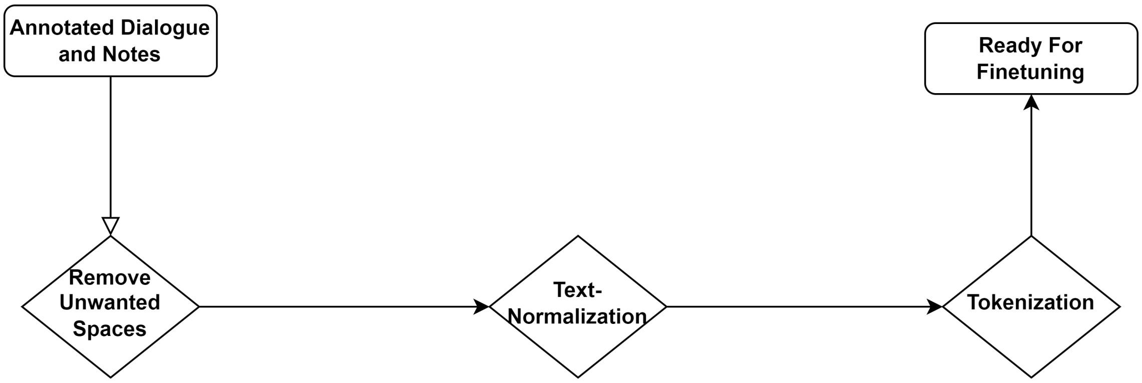


Figure 11.1: Preprocessing Steps for Preparing data to fine-tune

Removing Unnecessary Spaces. Our proposed dataset is suitable for clinical documentation. However, we had to remove certain spaces and line gaps for text processing and analysis to streamline the training.

Text-Normalization. According to [30], Abugida text in digital space contains a significant amount of errors and ambiguity. Bangla is no exception to this issue. That is why we have normalized our text with `BnUnicodeNormalizer`.

Tokenization. Tokenization is one of the crucial steps in pre-processing pipeline. As we have used mainly `HuggingFace` models. We have used `AutoModelTokenizer`.

11.4 System Information

Table 11.1: Brief Computational System information

Dataset	GPU	VRAM	Online Provider
BnClinical-Sum-Short	P100	16GB	Kaggle
BnClinical-Sum-Long	GTX A6000 Ada	48GB	vast.ai

Short Dataset Fine-tuning. We have used Kaggle's P100 GPU (16GB) for fine-tuning our datasets. As the token length of each pair of data was quite reasonable, we were able to use our model within the P100 GPU.

Long Dataset Fine-tuning. For fine-tuning, we have used one RTX A6000Ada GPU. As the input token length of long dataset is quite large, we had to incorporate a GPU with large VRAM (48 GB).

11.5 Performance Comparison

11.5.1 Evaluation Metric

For evaluating performance, we have used BERTScore[31]. It is a deep learning-based evaluation metric that uses contextual embedding, pairwise cosine similarity, and inverse document frequency. The overall workflow is like Figure 11.2.

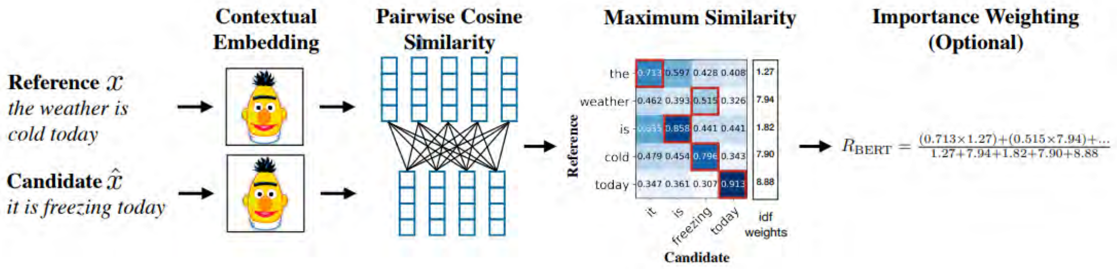


Figure 11.2: Workflow of BERTScore. It shows that how Reference x and Candidate $\hat{x}$ are evaluating using contextual embedding, pairwise cosine similarity, and inverse document frequency. [31]

Contextual Embedding. Contextual embedding actually provides how the overall reference text x and candidate text $\hat{x}$ are positioned in multidimensional tensors of a contextual embedding. By comparing both tensors the text similarities can be calculated.

Pairwise Cosine Similarity. To understand how those tensors are similar, it is essential to use certain metrics. At this stage, using Pairwise Cosine Similarity the similarity matrix is calculated. The formula for Pairwise Cosine Similarity is 11.1

$$\frac{x_i^\top \hat{x_j}}{||x_i|| \cdot ||\hat{x_j}||} \tag{11.1}$$

To calculate recall, precision, and F1 score Equation 11.2, 11.3, and 11.4 are used.

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x_j} \in \hat{x}} x_i^\top \hat{x_j} \tag{11.2}$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{x_i \in x} \max_{\hat{x_j} \in \hat{x}} x_i^\top \hat{x_j} \tag{11.3}$$

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \tag{11.4}$$

11.5.2 Short Dataset Fine-tuning

Table 11.2: Validation Set Performance on Short Data: BERT Score

Model Name	Precision	Recall	F1
CrossSum_Enhanced	0.7685	0.7181	0.7411
BanglaT5	0.6205	0.6606	0.6386
mT5 Multilingual XLSum	0.7657	0.7210	**0.7414**
mBART	0.7441	0.7362	0.7389
CrossSum	0.7683	0.7179	0.7409

Table 11.3: Test Set Performance on Short Data: BERT Score

Model Name	Precision	Recall	F1
mT5 multilingual XLSum	0.7710	0.7309	**0.7489**
CrossSum	0.7653	0.7201	0.7406
mBART	0.7504	0.7459	0.7470
CrossSum_Enhanced	0.7703	0.7244	0.7452
BanglaT5	0.6216	0.6700	0.6435

From Table 11.3 and 11.2, we can easily understands that top performing model for short dialogue to note task is mT5 Multilingual XLSum. It's F1 score is 0.7414. On the other hand, BanglaT5 gives us inferior results. Our hypothesis behind this results is that mT5 Multilingual XLSum pretrained on a sligthly recent data, so it have content of COVID-19 which slightly correlated with clinical Data.

11.5.3 Loss Curves For Short Dialogue Dataset Fine-tuning

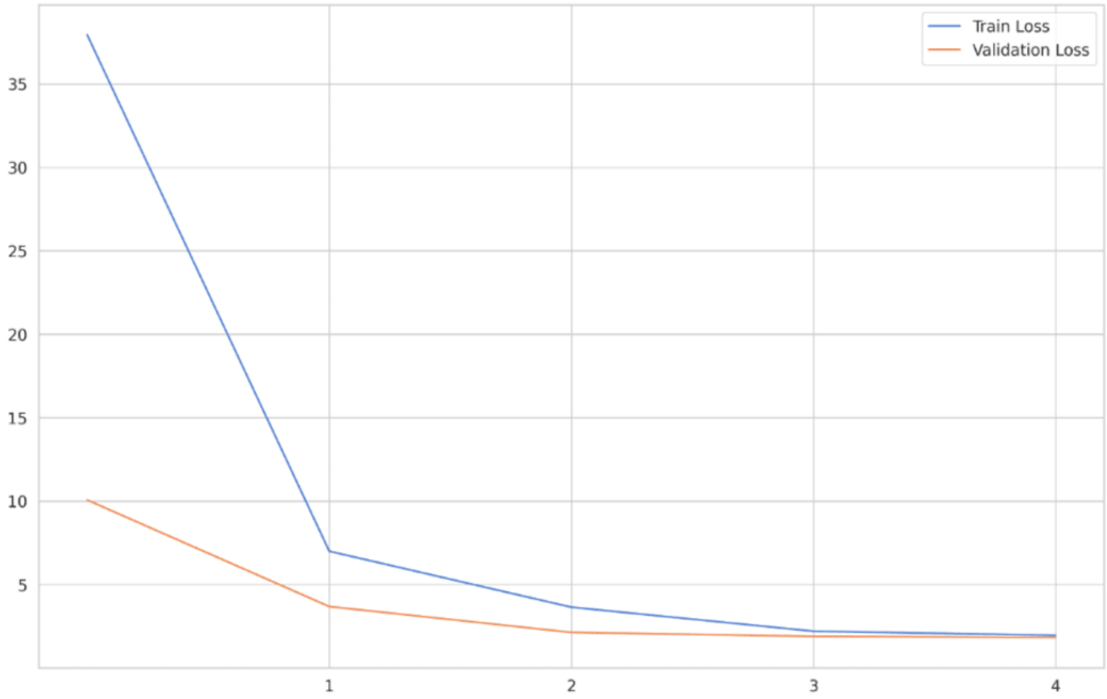


Figure 11.3: Loss Curve for BanglaT5. X-Axis contains an epoch count ranging from 0 to 4. Y-Axis represents a loss. Here, blue line represents training loss and orange line represents validation loss. For short dialogue generation.

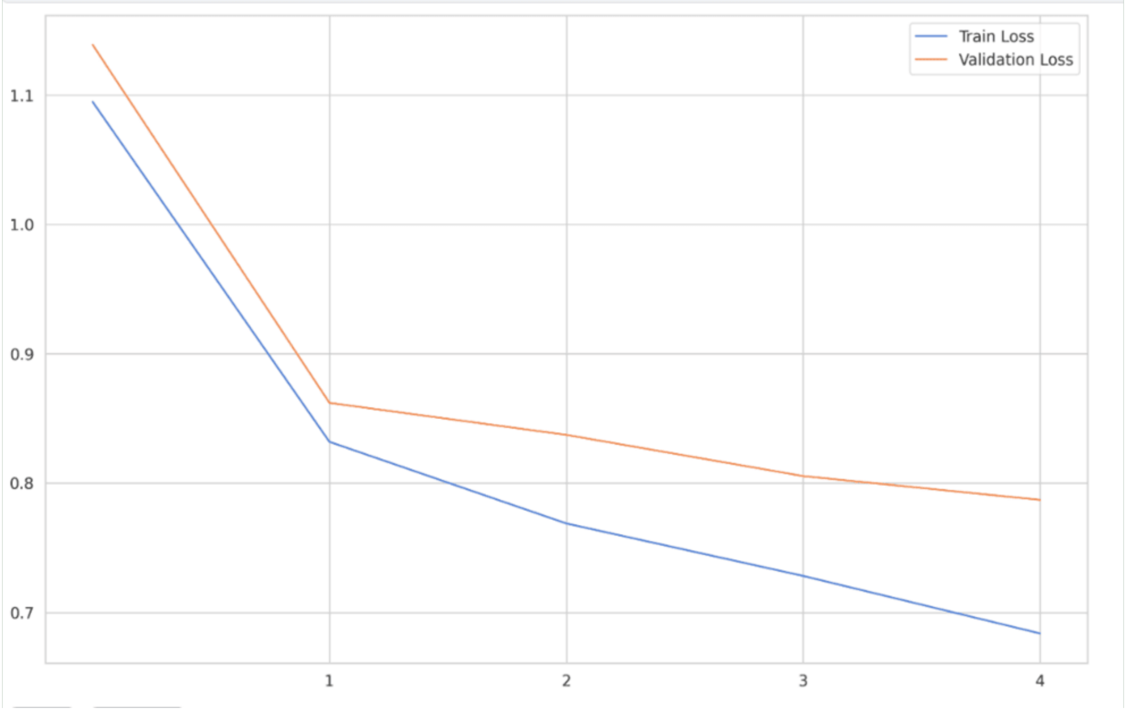


Figure 11.4: Loss Curve for CrossSum. X-Axis contains an epoch count ranging from 0 to 4. Y-Axis represents a loss. Here, blue line represents training loss and orange line represents validation loss. For short dialogue generation.

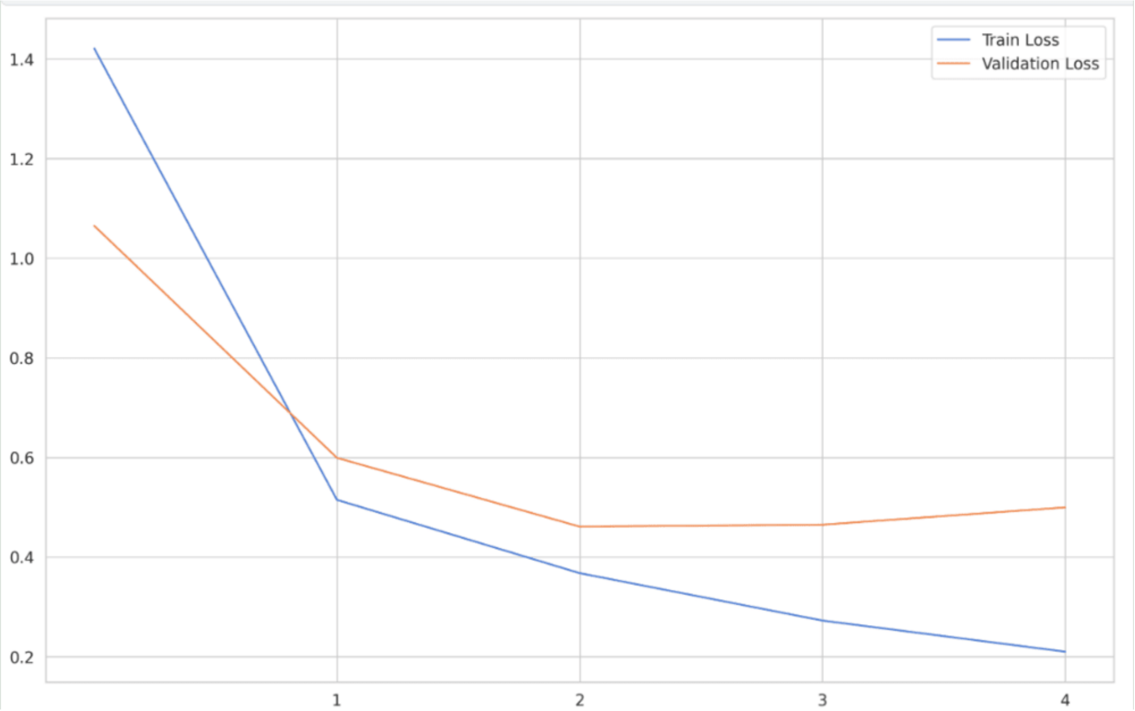


Figure 11.5: Loss Curve for mbart large. X-Axis contains an epoch count ranging from 0 to 4. Y-Axis represents a loss. Here, blue line represents training loss and orange line represents validation loss. For short dialogue generation.

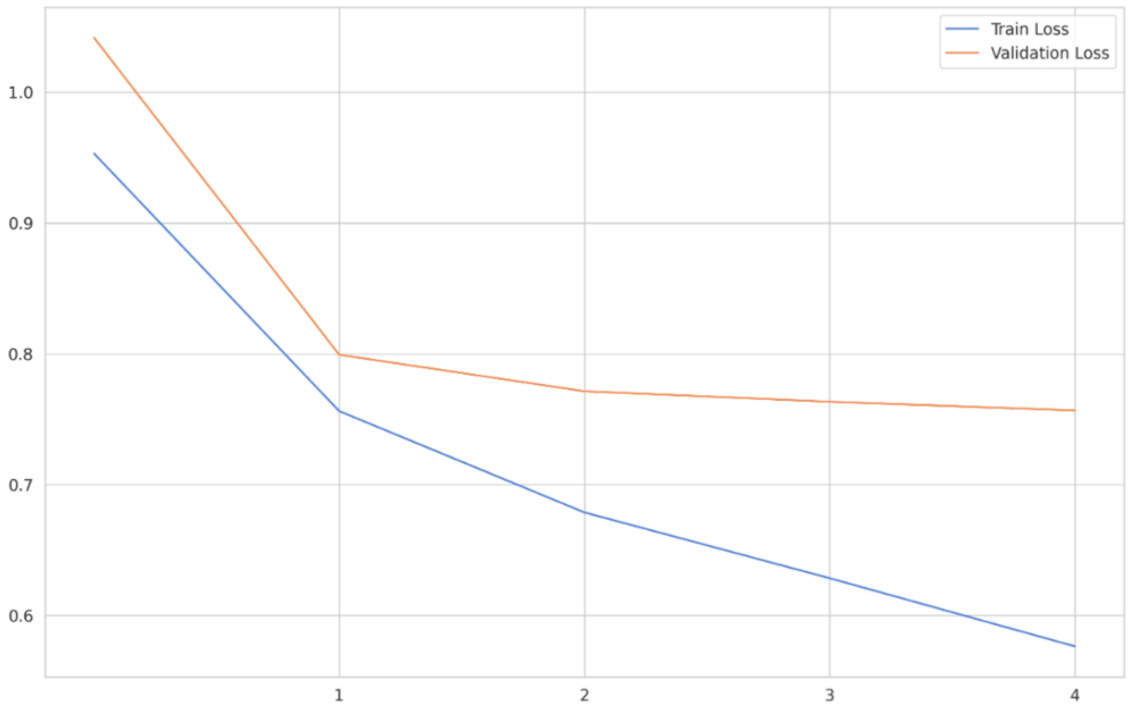


Figure 11.6: Loss Curve for CrossSum Enhanced. X-Axis contains an epoch count ranging from 0 to 4. Y-Axis represents a loss. Here, blue line represents training loss and orange line represents validation loss. For short dialogue generation.

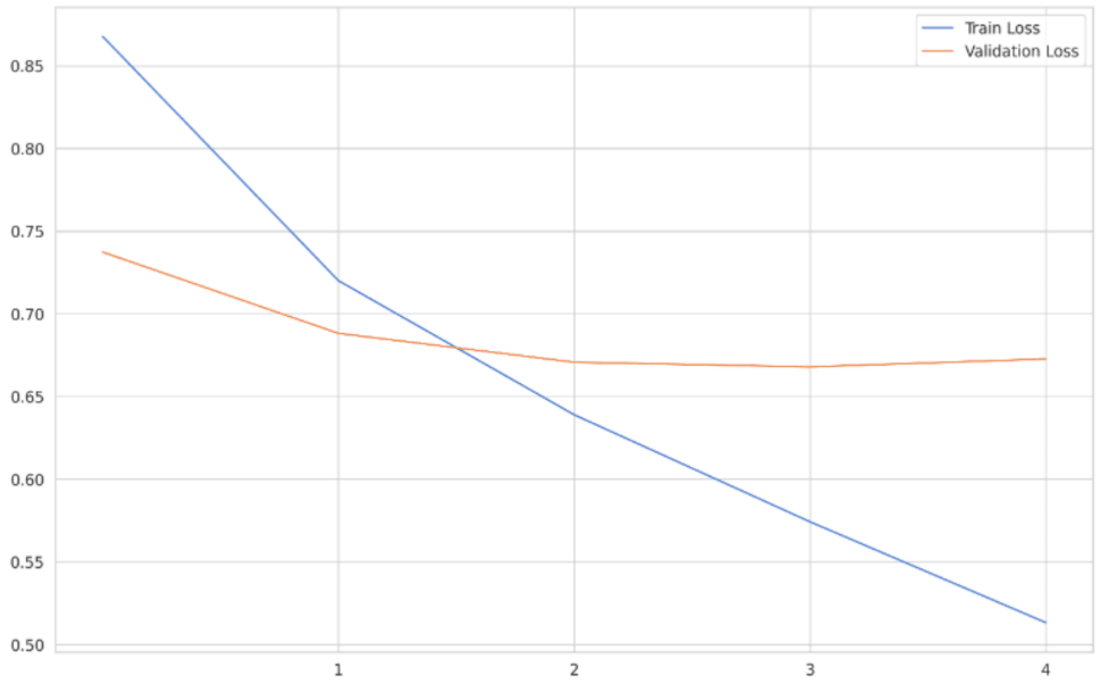


Figure 11.7: Loss Curve for Multilingual XLSum. X-Axis contains an epoch count ranging from 0 to 4. Y-Axis represents a loss. Here, blue line represents training loss and orange line represents validation loss. For short dialogue generation.

11.5.4 Long Dataset Fine-tuning

Table 11.4: Validation Set Performance on Long Data: BERT Score

Model Name	**Precision**	**Recall**	**F1**
mLongT5_base	0.6325	0.5803	0.6048
mT5 Multilingual XLSum	0.7591	0.7087	**0.7329**

Table 11.5: Test Set Performance on Long Data: BERT Score

Model Name	**Precision**	**Recall**	**F1**
mT5 Multilingual XLSum	0.7587	0.7037	**0.7299**
mLongT5_base	0.6471	0.5866	0.6150

For benchmark long dataset, we have restricts our experimentation to two model. First one is mLongT5 another is mT5 Multilingual XLSum. Reason behind, this restriction is that, other models are not good enough for handle long sequence generation tasks. For both Test and Validation, we have seen that mT5 Multilingual XLSum performs better. Results can be found in Table 11.5 and 11.4. This is actully a surprising results for us because, we expected a mLongT5 might produce better result due to it's capability of handling long sequences better than mT5 Multilingual XLSum. But our experimentation shows that mT5 Multilingual XLSum performed well.

11.5.5 Loss Curves For Long Dialogue Dataset Fine-tuning

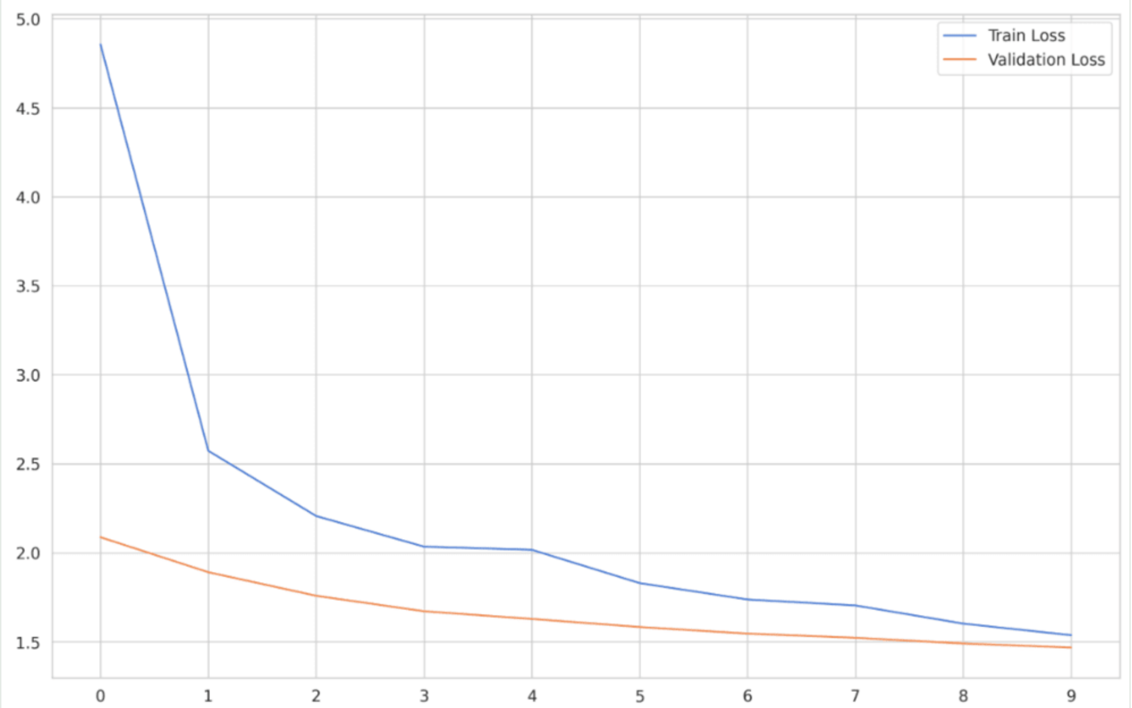


Figure 11.8: Loss Curve for mLongT5. X-Axis contains an epoch count ranging from 0 to 9. Y-Axis represents a loss. Here, blue line represents training loss and orange line represents validation loss. For long dialogue generation.

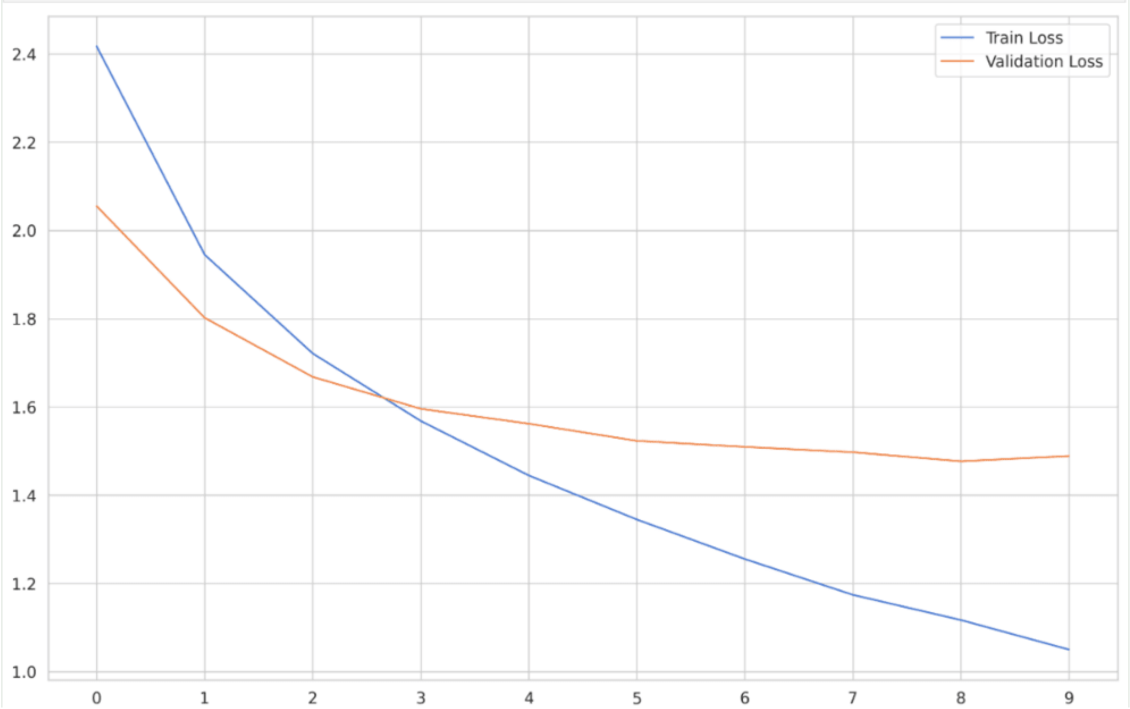


Figure 11.9: Loss Curve for Multilingual XLSum. X-Axis contains an epoch count ranging from 0 to 9. Y-Axis represents a loss. Here, blue line represents training loss and orange line represents validation loss. For long dialogue generation.

Chapter 12

Limitations

12.1 Time Comparison

From our observation in Table 12.1, we can see that the inference time varies a lot. Also, overall the model is not fast enough for real-world integration. In the case of long text generation, we can see that. It took around 50 sec to generate one output. So, there are a lot of room for model optimization. Also, we can see that mT5 Multilingual XLSum provides better inference time. Still, it is not enough for real world applications.

Table 12.1: Time Consumption of Models (Short Dataset)

Model	**Inference Time (sec/item)**	**Training Time (sec/epoch)**
CrossSum	8.43	404
CrossSum_Enhanced	9.065	415
mBART	17.76	513.8
mT5 Multilingual XLSum	**6.692**	550
BanglaT5	9.325	475

Table 12.2: Time Consumption of Models (Long Dataset)

Model	**Inference Time (sec/item)**	**Training Time (sec/epoch)**
mT5 Multilingual XLSum	**12.87**	134.8 (On A6000 Ada GPU)
mLongT5_base	50	186 (On A6000 Ada GPU)

12.2 Biasness in Dataset

Although, we have created a Bangla dataset. But it is based on an English dataset. That is why it encapsulates most of the details from the Western world. So, the dataset might be biased against Bangladesh. As the English dataset does not capture the demography of Bangladesh.

12.3 Hardware Limitation

One of the core problems in our work, we faced it hardware limitations. To train a model for long sequence generation, it is essential to have a high VRAM GPU. But currently available GPUs do not have that much VRAM. This situation makes the training process difficult.

Chapter 13

Future Work

13.1 Loss function in Text Summarization

We have observed that loss function does not correlated with generation quality. That's is why model training process might not be well. This problem can be divided into two segments, First of all there might be problems with loss function or evaluation metric. Both of those components of a model can be a way to explore.

13.2 Development of Corpus

We have a very limited amount of data in the Bangla Clinical NLP domain. During performance comparison, we have seen that Multilingual XLSum performs better. To our observation, we have seen that this model is trained with new article data initially and has access to COVID-19-related data. COVID-19-related data sort of gives insides of clinical data to a model. That is why we have to develop a corpus will aid this types of tasks.

13.3 Development of BnClinical-T5

Developing a pre-trained model might help to aid general Bangla clinical tasks. As for English, Clinical-T5 performs better than T5 for this type of task. So, the development of Bangla clinical T5 can help this problem to be solved.

13.4 Development of Pipeline

In this problem, we are dealing with long sequences. Also, long sequence models are not mature enough. That is why we can explore certain pipelines that can leverage short dialogues to generate sequence.

Chapter 14

Conclusion

To conclude, our ultimate objective was to convert the doctor-patient conversation into a Bengali medical report. To generate this dataset, we therefore needed to collaborate with various additional datasets. Approximately 1301 data sets have been utilized. For us, Google translator has made this seemingly straightforward. Nevertheless, we ran across a lot of issues with this translation, which we were able to resolve and provide a reliable dataset to find. The entire set of data that we have worked with was personally dealt with, and the source of the data was verified to assure its validity. We intended to make a difference in healthcare sector to improve our quality of life and save time. Our dataset will be very helpful to anyone working in this field in the future and will simplify their task.

Bibliography

[1] Asma Ben Abacha et al. "An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters." In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics.* Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2291–2302. DOI: 10.18653/v1/2023.eacl-main.168. URL: https://aclanthology.org/2023.eacl-main.168.

[2] Wen-wai Yim et al. "Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation." In: *Scientific Data* 10.1 (Sept. 2023), p. 586. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02487-3. URL: https://doi.org/10.1038/s41597-023-02487-3.

[3] Gagandeep Singh et al. "Large Scale Sequence-to-Sequence Models for Clinical Note Generation from Patient-Doctor Conversations." In: *Proceedings of the 5th Clinical Natural Language Processing Workshop.* Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 138–143. DOI: 10.18653/v1/2023.clinicalnlp-1.18. URL: https://aclanthology.org/2023.clinicalnlp-1.18.

[4] Kadir Bulut Ozler and Steven Bethard. "clulab at MEDIQA-Chat 2023: Summarization and classification of medical dialogues." In: *Proceedings of the 5th Clinical Natural Language Processing Workshop.* Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 144–149. DOI: 10.18653/v1/2023.clinicalnlp-1.19. URL: https://aclanthology.org/2023.clinicalnlp-1.19.

[5] Ashwyn Sharma, David Feldman, and Aneesh Jain. "Team Cadence at MEDIQA-Chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models." In: *Proceedings of the 5th Clinical Natural Language Processing Workshop.* Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 228–235. DOI: 10.18653/v1/2023.clinicalnlp-1.28. URL: https://aclanthology.org/2023.clinicalnlp-1.28.

[6] Junda Wang et al. "UMASS_BioNLP at MEDIQA-Chat 2023: Can LLMs generate high-quality synthetic note-oriented doctor-patient conversations?" In: *Proceedings of the 5th Clinical Natural Language Processing Workshop.* Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 460–471. DOI: 10.18653/v1/2023.clinicalnlp-1.49. URL: https://aclanthology.org/2023.clinicalnlp-1.49.

[7] Kunal Suri, Saumajit Saha, and Atul Singh. "HealthMavericks@MEDIQA-Chat 2023: Benchmarking different Transformer based models for Clinical Dialogue Summarization." In: *Proceedings of the 5th Clinical Natural Language Processing Workshop.* Toronto, Canada: Association for Computational Lin-

guistics, July 2023, pp. 472–489. DOI: 10.18653/v1/2023.clinicalnlp-1.50. URL: https://aclanthology.org/2023.clinicalnlp-1.50.

[8] Xiangru Tang et al. "GersteinLab at MEDIQA-Chat 2023: Clinical Note Summarization from Doctor-Patient Conversations through Fine-tuning and In-context Learning." In: *Proceedings of the 5th Clinical Natural Language Processing Workshop.* Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 546–554. DOI: 10.18653/v1/2023.clinicalnlp-1.58. URL: https://aclanthology.org/2023.clinicalnlp-1.58.

[9] Boya Zhang, Rahul Mishra, and Douglas Teodoro. "DS4DH at MEDIQA-Chat 2023: Leveraging SVM and GPT-3 Prompt Engineering for Medical Dialogue Classification and Summarization." In: *Proceedings of the 5th Clinical Natural Language Processing Workshop.* Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 536–545. DOI: 10.18653/v1/2023.clinicalnlp-1.57. URL: https://aclanthology.org/2023.clinicalnlp-1.57.

[10] Kirill Milintsevich and Navneet Agarwal. "Calvados at MEDIQA-Chat 2023: Improving Clinical Note Generation with Multi-Task Instruction Finetuning." In: *Proceedings of the 5th Clinical Natural Language Processing Workshop.* Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 529–535. DOI: 10.18653/v1/2023.clinicalnlp-1.56. URL: https://aclanthology.org/2023.clinicalnlp-1.56.

[11] Dhananjay Srivastava. "IUTEAM1 at MEDIQA-Chat 2023: Is simple fine tuning effective for multi layer summarization of clinical conversations?" In: *Proceedings of the 5th Clinical Natural Language Processing Workshop.* Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 519–523. DOI: 10.18653/v1/2023.clinicalnlp-1.54. URL: https://aclanthology.org/2023.clinicalnlp-1.54.

[12] Yash Mathur et al. "SummQA at MEDIQA-Chat 2023: In-Context Learning with GPT-4 for Medical Summarization." In: *Proceedings of the 5th Clinical Natural Language Processing Workshop.* Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 490–502. DOI: 10.18653/v1/2023.clinicalnlp-1.51. URL: https://aclanthology.org/2023.clinicalnlp-1.51.

[13] Alex Papadopoulos Korfiatis et al. "PriMock57: A Dataset Of Primary Care Mock Consultations." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 588–598. DOI: 10.18653/v1/2022.acl-short.65. URL: https://aclanthology.org/2022.acl-short.65.

[14] Longxiang Zhang et al. "Leveraging Pretrained Models for Automatic Summarization of Doctor-Patient Conversations." In: *Findings of the Association for Computational Linguistics: EMNLP 2021.* Ed. by Marie-Francine Moens et al. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3693–3712. DOI: 10.18653/v1/2021.findings-emnlp.313. URL: https://aclanthology.org/2021.findings-emnlp.313.

[15] Kundan Krishna et al. "Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 4958–4972. DOI: 10.18653/v1/2021.acl-long.384. URL: https://aclanthology.org/2021.acl-long.384.

[16] Wen-wai Yim and Meliha Yetisgen. "Towards Automating Medical Scribing : Clinic Visit Dialogue2Note Sentence Alignment and Snippet Summarization." In: *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*. Ed. by Chaitanya Shivade et al. Online: Association for Computational Linguistics, June 2021, pp. 10–20. DOI: 10.18653/v1/2021.nlpmc-1.2. URL: https://aclanthology.org/2021.nlpmc-1.2.

[17] Gregory Finley et al. "From dictations to clinical reports using machine translation." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. Ed. by Srinivas Bangalore, Jennifer Chu-Carroll, and Yunyao Li. New Orleans - Louisiana: Association for Computational Linguistics, June 2018, pp. 121–128. DOI: 10.18653/v1/N18-3015. URL: https://aclanthology.org/N18-3015.

[18] Seppo Enarvi et al. "Generating Medical Reports from Patient-Doctor Conversations Using Sequence-to-Sequence Models." In: *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*. Ed. by Parminder Bhatia et al. Online: Association for Computational Linguistics, July 2020, pp. 22–30. DOI: 10.18653/v1/2020.nlpmc-1.4. URL: https://aclanthology.org/2020.nlpmc-1.4.

[19] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 `[cs.CL]`.

[20] Jay Alammar. "The illustrated transformer." In: *The Illustrated Transformer–Jay Alammar–Visualizing Machine Learning One Concept at a Time* 27 (2018).

[21] Yuqing Tang et al. "Multilingual Translation from Denoising Pre-Training." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 3450–3466. DOI: 10.18653/v1/2021.findings-acl.304. URL: https://aclanthology.org/2021.findings-acl.304.

[22] Yinhan Liu et al. "Multilingual Denoising Pre-training for Neural Machine Translation." In: *Transactions of the Association for Computational Linguistics* 8 (Nov. 2020), pp. 726–742. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00343. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00343/1923401/tacl_a_00343.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00343.

[23] Tahmid Hasan et al. "CrossSum: Beyond English-Centric Cross-Lingual Abstractive Text Summarization for 1500+ Language Pairs." In: *CoRR* abs/2112.08804 (2021). arXiv: 2112.08804. URL: https://arxiv.org/abs/2112.08804.

[24] Linting Xue et al. *mT5: A massively multilingual pre-trained text-to-text transformer*. 2021. arXiv: 2010.11934 `[cs.CL]`.

[25] Yi Tay et al. *UL2: Unifying Language Learning Paradigms.* 2023. arXiv: 2205.05131 [cs.CL].

[26] Mandy Guo et al. *LongT5: Efficient Text-To-Text Transformer for Long Sequences.* 2022. arXiv: 2112.07916 [cs.CL].

[27] Tahmid Hasan et al. "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.* Online: Association for Computational Linguistics, Aug. 2021, pp. 4693–4703. URL: https://aclanthology.org/2021.findings-acl.413.

[28] Abhik Bhattacharjee et al. "BanglaNLG and BanglaT5: Benchmarks and Resources for Evaluating Low-Resource Natural Language Generation in Bangla." In: *Findings of the Association for Computational Linguistics: EACL 2023.* Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 726–735. DOI: 10.18653/v1/2023.findings-eacl.54. URL: https://aclanthology.org/2023.findings-eacl.54.

[29] Abhik Bhattacharjee et al. *BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla.* 2022. arXiv: 2101.00204 [cs.CL].

[30] Nazmuddoha Ansary et al. *Abugida Normalizer and Parser for Unicode texts.* 2023. arXiv: 2306.01743 [cs.CL].

[31] Tianyi Zhang et al. *BERTScore: Evaluating Text Generation with BERT.* 2020. arXiv: 1904.09675 [cs.CL].