

**Zero-Shot Evaluation of Large Language Models on Medical Query
Summarization Tasks**

Abrar Chowdhury

190042140

Md. Sadman Sakib

190042111

Farhan Shahriar Fahim

190042127

Department of Computer Science and Engineering

Islamic University of Technology

June, 2024

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Abrar Chowdhury, Md. Sadman Sakib,** and **Farhan Shahriar Fahim** under the supervision of **Dr. Md Moniruzzaman**, Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Dr. Md Moniruzzaman

Assistant Professor

Department of Computer Science and Engineering

Islamic University of Technology (IUT)

Date: June 05, 2024

Abrar Chowdhury

Student ID: 190042140

Date: June 05, 2024

Md. Sadman Sakib

Student ID: 190042111

Date: June 05, 2024

Farhan Shahriar Fahim

Student ID: 190042127

Date: June 05, 2024

Dedicated to Dr. Md Moniruzzaman, whose expert guidance, insightful feedback, and steadfast support have been crucial to the completion of this work. His dedication to excellence and commitment to our academic and professional development have been invaluable. We deeply grateful for his mentorship and encouragement throughout this journey.

Contents

1	Introduction	1
1.1	Large Language Models	1
1.2	Motivations and Scope	6
1.3	Problem Statement	7
1.4	Research Challenges	7
1.5	Contribution	8
1.6	Organization	9
2	Related Works	11
3	Proposed Methodology	15
3.1	Dataset Selection	15
3.2	Model Selection and Zero-Shot Learning Approach	16
3.3	Benchmarking and Evaluation Framework	16
3.3.1	Inference Pipeline for LLM Evaluation	17
3.3.2	Post-Processing of Model Outputs:	17
3.3.3	Evaluation Metrics for Response Quality	17
3.4	Analysis and Interpretation of Results	19
4	Results and Discussion	20
4.1	Analysis and Comparison	22
5	Conclusion	24
5.1	Conclusion	24
5.2	Future Work	25
5.2.1	Enhancement of Model Robustness	25
5.2.2	Integration of Multimodal Approaches	25
5.2.3	Exploration of Zero-shot Learning Strategies	25
5.2.4	Evaluation on Larger and Diverse Datasets	25

5.2.5	User-Centric Evaluation and Deployment	26
5.2.6	Ethical and Societal Implications	26

References		27
-------------------	--	-----------

List of Tables

4.1	Samples of summaries generated by the benchmark models	21
4.2	Automatic evaluation results of BanglaCHQ-Summ	23
4.3	BLEU Score evaluation results of BanglaCHQ-Summ	23

List of Abbreviations

CNN	Convolutional Neural Network
PIP	PIP Installs Packages
TikZ	TikZ ist kein Zeichenprogramm
WIKI	What I Know Is
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
BERTscore	BERT-based Score
CHQs	Consumer Health Questions
COVID-19	Coronavirus Disease 2019
DL	Deep Learning
EM	Exact Match
F1	F1 Score
GPT	Generative Pre-trained Transformer
LLMs	Large Language Models
ML	Machine Learning
NLP	Natural Language Processing
NLU	Natural Language Understanding
NLG	Natural Language Generation
QA	Question Answering
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
T5	Text-To-Text Transfer Transformer

Acknowledgement

I would like to express my sincere gratitude to Dr. Md Moniruzzaman for his invaluable guidance, support, and encouragement throughout the duration of this project. His insightful feedback and constructive suggestions greatly contributed to the quality and completion of this report.

I would like to also convey my heartfelt appreciation for Md Tahmid Rahman Laskar, for he voluntarily took time out of his busy schedule to offer us counsel, idea and insight.

I am deeply grateful to my family and friends for their unwavering support, patience, and encouragement. Their understanding and motivation have been a source of strength throughout my academic journey.

Determination and the mindset to never give up no matter how much has been lost are two virtues that brought this work to life. I think from here on out I will appreciate them more than I have in the first 3 years of my undergrad life.

I thank you all for your contributions and support.

Abstract

Online health consultations are becoming increasingly popular as a way for patients to discuss their medical health inquiries. In Bangladesh, patients are also using online health care solutions and thus providing medical queries in Bangla language. The COVID-19 pandemic has accelerated the use of these platforms, leading to a significant influx of questions and placing a heavy burden on the limited number of healthcare professionals available to respond. Text summarization offers a promising solution by condensing Bangla medical queries to highlight only the essential information needed for answers. This not only reduces the time healthcare professionals spend parsing unnecessary details but also serves as a crucial step toward developing automated medical question-answering systems. This research presents a comprehensive zero-shot evaluation of several state-of-the-art Bangla and multilingual text generation models on the task of summarizing Bangla Consumer Health Questions (CHQs). The models we evaluated include BanglaT5, mT5, GPT-3.5, and GPT-4. The evaluation was conducted using ‘BanglaCHQ-Summ,’ which is currently the only available dataset specifically designed for summarizing Bangla CHQs, comprising 2350 pairs of questions and their corresponding summaries. The study aimed to determine which model performs best in terms of accurately and concisely summarizing Bangla medical queries. Among the models tested, GPT-4 demonstrated superior performance, achieving a BERTScore of 90.25%.

Chapter 1

Introduction

This section offers a detailed overview of the research, highlighting its motivation and scope, problem statement, contributions and challenges. The rapid advancement of digital technologies has reformed healthcare delivery, with online health consultations becoming a popular platform for Bangla-speaking patients to seek medical advice, a trend accelerated by the COVID-19 pandemic. Healthcare professionals are experiencing a sense of being overwhelmed by the sudden increase, which has drawn attention to the necessity for efficient ways to manage large volumes of inquiries. Abstractive text summarising is a method that can help medical specialists save time by evaluating material and generating concise summaries of medical questions in the Bangla language. This will enable the creation of automated systems capable of responding to medical inquiries. Due to limited resources in natural language processing, summarising Bangla Consumer Health Questions (CHQs) is a difficult task. This study examines the 'BanglaCHQ-Summ' dataset, which consists of 2,350 question-summary pairs, and compares it to the 'BanglaT5, mT5, GPT-3.5, and GPT-4' datasets, which are the only datasets specifically designed for summarising Bangla CHQs. According to our research, GPT-4 shown significant potential in enhancing the efficiency of online health consultations for those who speak Bangla. This was evidenced by its remarkable BERTScore of 90.25%. This study establishes the foundation for future investigations in the field by presenting the dataset and doing a comprehensive analysis of performance.

1.1 Large Language Models

The science of natural language processing (NLP) is continuously advancing, and one of the most efficient tools for understanding and generating human-like writing is the large language model (LLM). Assessing the performance of these models on specific tasks, par-

ticularly in specialised fields such as medical question summarising, remains challenging. Large Language Models (LLMs) utilise extensive datasets and intricate neural network architectures to comprehend and generate text that has a genuine tone. The GPT series developed by OpenAI, BERT created by Google, and LLaMA developed by Meta are all instances of Language Model Libraries (LLMs).

LLMs operate through a combination of neural networks, extensive training data, and statistical methods. Their workflow can be summarized as follows:

- **Training Data**

- **Massive Datasets:** LLMs are trained on vast text corpora encompassing books, articles, websites, and various other written materials. For instance, GPT-4 was trained on diverse datasets containing billions of words, covering a wide range of topics and styles.
- **Diverse Sources:** The training data spans multiple domains, allowing LLMs to acquire a broad understanding of different contexts and linguistic nuances. This diversity ensures that the models can handle a wide array of queries and generate contextually relevant responses.

- **Neural Architecture**

- **Transformers:** The Transformer architecture, which allowed fast handling of huge datasets and revolutionised natural language processing, is mostly used by LLMs. The Transformer's multi-layer architecture uses self-attention processes to rank the significance of individual words inside a given phrase.
- **Layers:** Each of the many levels that make up an LLM (for example, GPT-4 has 96 layers) contains many attention heads that pick up on distinct facets of the input. The model is able to construct sophisticated language and context representations through these levels.

- **Training Process**

- **Pretraining:** During the pretraining phase, LLMs learn language patterns by predicting the next word in a sentence given the preceding words. This phase involves training the model on large-scale text data to learn grammatical structures, word associations, and contextual relationships.
- **Fine-Tuning:** After pretraining, the model undergoes fine-tuning on specific tasks or datasets to enhance its performance in particular areas. For example,

fine-tuning on medical texts enables the model to better handle medical queries and terminology.

- **Self-Attention Mechanism**

- **Contextual Understanding:** The self-attention mechanism allows LLMs to focus on different parts of a text to understand its context. This mechanism assigns different weights to words based on their relevance, helping the model grasp the meaning of sentences in a nuanced manner.
- **Bidirectionality:** In models like BERT, the bidirectional approach reads text from both directions (left-to-right and right-to-left), providing a deeper understanding of context by considering the entire sentence structure.

- **Tokenization**

- **Splitting Text:** LLMs convert text into tokens, which are smaller units (often words or subwords). Tokenization enables the model to handle text in a structured format and facilitates the analysis of linguistic patterns.
- **Embedding:** Tokens capture word-to-token semantic links by being represented as vectors in a three-dimensional space. Thanks to these embeddings, the model can grasp the interconnectedness of words and their meanings.

- **Generating Text**

- **Decoding:** When generating text, the model predicts one token at a time, selecting the most probable next token based on the given context. This process involves using algorithms like beam search or sampling to produce coherent and contextually relevant outputs.
- **Context Window:** LLMs have a context window (e.g., 8192 tokens in GPT-4-turbo), which limits the amount of text they can consider at once. This window size affects how well the model can understand and generate long passages of text, as it determines the amount of context available for each prediction.

- **Adaptation and Fine-Tuning**

- **Task-Specific Adjustments:** LLMs can be adapted to specific tasks (e.g., translation, summarization) by fine-tuning on task-specific datasets. This process adjusts the model parameters to better suit the desired application, enhancing its performance on specialized queries.
- **Prompt Engineering:** Users can influence LLMs' responses by crafting spe-

cific input prompts. Prompt engineering involves designing prompts that guide the model to generate outputs aligned with the intended purpose, such as answering medical questions accurately.

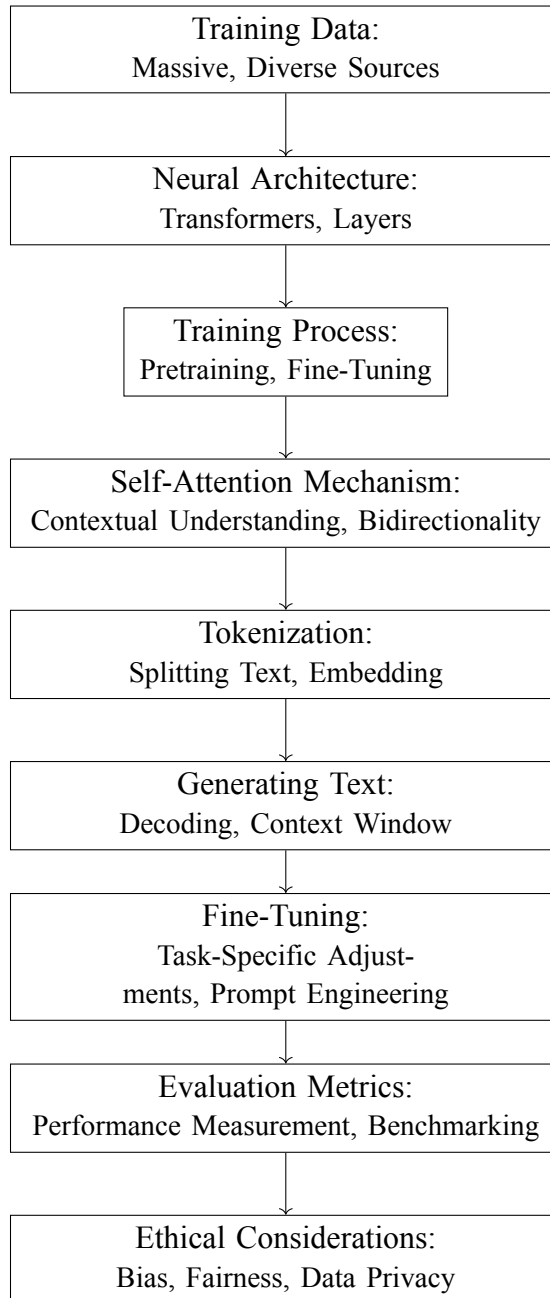


Figure 1.1: Workflow of Large Language Models (LLMs).

- **Evaluation and Metrics**

- **Performance Measurement:** LLMs are evaluated using metrics like perplexity (which measures prediction accuracy) and BLEU scores (which assess the quality of generated text against reference texts). These metrics help quantify the effectiveness of the model in understanding and generating language.
- **Benchmarking:** Various benchmarks and datasets are used to assess how well LLMs perform across different tasks and domains. These benchmarks provide standardized evaluation frameworks for comparing different models and measuring their capabilities.

- **Ethical Considerations**

- **Bias and Fairness:** LLMs can inherit biases present in the training data. Addressing these biases and ensuring fairness in AI-generated content is a critical area of ongoing research, as biased outputs can impact decision-making and perpetuate stereotypes.
- **Data Privacy:** Handling sensitive information, especially in domains like healthcare, requires careful consideration to ensure data privacy and compliance with regulations. Ensuring that AI systems protect user data and adhere to ethical standards is essential for their responsible deployment.

- **Real-World Applications:**

The ability to accurately summarize medical queries holds profound implications across the healthcare spectrum, benefiting professionals, researchers, and patients. Efficient summarization aids in rapid information retrieval and decision-making by condensing complex medical information into coherent summaries, thus allowing healthcare providers to quickly grasp essential details and respond promptly in clinical settings.

For researchers, summarization streamlines the synthesis of vast amounts of medical knowledge, facilitating the identification of trends and accelerating the pace of innovation by distilling critical insights from the literature and patient data. For patients, summarized information enhances comprehension, enabling them to understand medical advice and treatment plans more clearly, which fosters better adherence and engagement with their care.

Moreover, summarization bridges communication gaps, particularly for those with limited health literacy or non-native language speakers, by presenting medical content in an accessible manner. On a systemic level, summarization improves health-

care outcomes by optimizing the flow of information, enhancing interoperability of electronic health records (EHRs), and contributing to more efficient data management and public health surveillance. Ultimately, the integration of advanced summarization techniques in healthcare supports improved decision-making, patient satisfaction, and the overall quality of care.

- **Key Terms and Concepts:**

- Large Language Models (LLMs) - Machine learning algorithms that have been trained on massive volumes of text data are able to comprehend and produce language that is very similar to human speech.
- Medical Query Summarization - The process of extracting key information from medical queries or documents to provide concise and relevant summaries.
- Zero-Shot Learning - Using AI to perform tasks without prior specific task-related training, based on generalized understanding.
- Natural Language Processing (NLP) - The science and technology that enables computers to comprehend, imitate, and process human speech.
- Prompt Engineering - Designing input prompts to elicit specific responses from AI models.
- Ethical AI in Healthcare - Addressing ethical considerations in the deployment of AI systems in medical environments.
- Knowledge Transfer - The ability of AI models to apply information learned from one domain to another, such as from general language to medical queries.
- Query Disambiguation - The AI's capability to resolve ambiguities in medical queries for precise understanding and response.

1.2 Motivations and Scope

Access to healthcare information and resources is a fundamental human right, yet for over 230 million Bengali-speaking individuals worldwide, this remains a significant challenge. The healthcare sector is particularly affected by the limitations of low-resource languages like Bengali, where the availability of medical subdomains is scarce. This research is motivated by the urgent need to bridge this gap in healthcare accessibility. By focusing on the Bengali language, we aim to provide more inclusive healthcare information, thereby contributing to the global effort of reducing disparities in healthcare access.

1.3 Problem Statement

Due to the rapid advancement of natural language processing (NLP) technology, more sophisticated models for text generation have surfaced. Bengali and other low-resource languages have not had their effectiveness on these models tested, which is particularly problematic in highly specialist domains such as healthcare. This research aims to address that gap by providing an in-depth analysis of current practice in text generation models for Bengali and other languages. The primary objective is to identify the most effective model for Bengali medical query summarising.

The research seeks to answer the following questions:

- **RQ1:** How do state-of-the-art language models (BanglaT5, mT5, GPT-3.5, and GPT-4) perform in summarizing Bengali medical queries?
- **RQ2:** How well do the models generalize to unseen medical subdomains or complex medical concepts not explicitly present in the training data?
- **RQ3:** How effective is the 'BanglaCHQ-Summ' dataset in training and evaluating models for Bengali medical query summarization?

In order to answer these problems, this study will compare and contrast the generalizability and performance of top natural language processing models on Bengali healthcare inquiries. The results will help direct future research and development in the area and add to our general knowledge of how well these models work in linguistic contexts with limited resources.

1.4 Research Challenges

Conducting a zero-shot evaluation on a Bengali medical health queries dataset presents several significant challenges, each of which impacts the effectiveness and efficiency of the research process. These challenges can be broadly categorized into linguistic, computational, and data-related issues.

Linguistic Challenges

Understanding Bengali syntax and morphology is inherently challenging due to the complexity and richness of the language. Bengali, like many other low-resource languages, has intricate grammatical structures and a diverse vocabulary that can pose difficulties for language models predominantly trained on high-resource languages. These models often

struggle to accurately capture the nuances of Bengali, leading to suboptimal performance in tasks such as text summarization and query understanding.

Computational Challenges

Training and evaluating large-scale language models require substantial computational resources. High-performance computing infrastructure is essential to handle the extensive data processing and model training involved in this research. The computational power needed to fine-tune and test models such as BanglaT5, mT5, GPT-3.5, and GPT-4 on the Bengali medical queries dataset is considerable. Ensuring that the models perform efficiently under these conditions is a critical aspect of the research.

Data-Related Challenges

Accessing high-quality Bengali medical query datasets and pre-trained language models tailored specifically to the Bengali language is another significant hurdle. The availability of comprehensive and annotated medical datasets in Bengali is limited, which restricts the training and evaluation of models. Additionally, the lack of pre-trained language models specifically designed for Bengali exacerbates this issue, as general multilingual models may not perform optimally on Bengali-specific tasks. Overcoming these data limitations is crucial for the success of the research.

In summary, addressing these challenges is vital for advancing the performance of NLP models in low-resource languages like Bengali. The research aims to contribute to the broader field by highlighting these issues and proposing solutions to enhance model effectiveness in handling Bengali medical queries.

1.5 Contribution

This study makes significant contributions to the field of natural language processing (NLP) with a specific focus on Bengali medical query summarization, an area that is both critical and underexplored in low-resource language settings. Our key contributions are as follows:

- **Comprehensive Evaluation Across Multiple Models:** We conducted a rigorous zero-shot evaluation of four advanced language models—BanglaT5, mT5, GPT-3.5, and GPT-4—on a dataset of Bengali medical health queries. This evaluation aimed to benchmark the performance of these models in summarizing medical queries, highlighting their strengths and limitations in processing Bengali text.

- **Showcasing Multilingual Model Potentials:** The study underscores the potential of multilingual models for addressing the challenges inherent in low-resource languages. By including both multilingual models (mT5, GPT-3.5, GPT-4) and a model specifically trained on Bengali (BanglaT5), we provide insights into how models with diverse training backgrounds handle the nuances of Bengali medical terminology and context.
- **Valuable Data for Low-Resource Language Processing:** Our findings offer valuable data on the performance of pre-trained models in a low-resource setting like Bengali. The comparative analysis sheds light on the efficacy of transfer learning and model adaptability, providing a foundation for future research and development in NLP for Bengali and other low-resource languages.
- **Methodological Contributions:** The study also contributes methodologically by employing a zero-shot evaluation approach. This method allows for assessing the models' inherent capabilities without additional fine-tuning on domain-specific data, thus offering a clear view of their baseline performance and adaptability.
- **Implications for Healthcare and NLP Research:** Our study has direct consequences for healthcare by concentrating on medical queries. Its primary goal is to improve access to medical information and help in Bengali-speaking countries. Better natural language processing (NLP) tools and resources for additional languages with less resources can be informed by the findings of this evaluation.

In conclusion, this work sheds light on the performance of multilingual and low-resource language models in medical query summarization, which opens up new possibilities for advancements and uses in this vital field.

1.6 Organization

The remainder of this thesis is organized into four main chapters, each designed to seamlessly guide the reader from the introduction to the conclusion, reflecting a logical flow of information aligned with our research objectives.

In **”Related Works”** chapter, we provide a comprehensive review of the existing literature relevant to Bengali language models and text summarization tasks. We discuss the advancements in natural language processing, focusing on previous studies related to medical query summarization and the development of Bengali language resources. This chapter sets the foundation for our research by highlighting the gaps and limitations in the current state of the art, thereby justifying the need for our study.

The next chapter details the methodology employed in our research. We describe the data preparation and preprocessing steps, including the construction and characteristics of the BanglaCHQ-Summ dataset. This chapter also outlines the models evaluated in our study, including BanglaT5, mT5, GPT-3.5, and GPT-4. We explain the experimental setup, the training and evaluation procedures, and the metrics used to assess model performance. This methodological framework ensures the reproducibility and robustness of our findings.

In **”Results and Discussion”** chapter, we present and analyze the results of our experiments. We compare the performance of the evaluated models based on metrics such as ROUGE and BERTScore, providing a detailed discussion of the findings. The chapter interprets the results in the context of the research objectives, explaining how each model performs in summarizing Bengali medical queries. We also discuss the implications of our findings for the field of NLP and healthcare, emphasizing the practical applications and potential impact of efficient summarization tools.

The final chapter concludes our work by summarizing the key contributions of our research, reflecting on the significance of our findings and their implications for future research. We restate the main research objectives and questions, providing a concise summary of the insights gained. The chapter acknowledges the limitations of our study and suggests potential directions for future exploration, such as refining model architectures and expanding evaluation datasets. We conclude with reflections on the broader impact of our work on improving healthcare accessibility and efficiency through advanced NLP techniques.

Chapter 2

Related Works

This study's overarching goal is to assess how well different benchmark models do when challenged with Bengali-language text summarising tasks. The 'BanglaCHQ-Summ' dataset, created specifically for assessment and benchmarking, will be used in the evaluation. Using a dataset of Bengali medical health questions across four distinct models, the study will perform a zero-shot assessment. Insights gained from this study about the strengths and weaknesses of existing text summarization models for low-resource languages will help direct future work in natural language processing that caters to Bengali and other marginalised languages.

A large number of work has been devoted to automated text summarization tasks for exclusive language like English (Gambhir and Gupta [5], Kumar, Kaur, and Kaur [11], and Mridha, Lima, Nur, *et al.* [13]). Although not as English, a decent amount of work has been dedicated to automated text summarization tasks for low resource language like Bangla (Haque, Pervin, Hossain, *et al.* [6], Sarkar and Hossen [15], and Talukder, Abujar, Masum, *et al.* [17]). Here they provided overview of various methods applied to Bangla text summarization, highlighting the challenges and techniques specific to the Bangla language. We explored and found some works on Bangla text summarization techniques. The work in Akter, Asa, Uddin, *et al.* [2] explores extractive summarization techniques for Bangla documents, employing sentence scoring and clustering methods. Given the exponential growth of textual data as a result of web-based applications, there is an urgent demand for automated solutions that may speed the summary process, saving time and resources. The most recent technique is abstractive text summarization. In answer to this need, Hayat, Das, and Hoque [7] presents an automated abstractive text summarising method for Bengali that uses transformer-based models.

As we are working on the medical domain, text summarization tasks are done already

for medical documents and prescriptions (Moradi and Ghadiri [12] and Sarker, Yang, Algaradi, *et al.* [16]). In Adib and Alam [1], the first long and short Bangla Clinical Dialogue to Note Summarization dataset of total 1,701 dialogue pairs and associated clinical note, BnClinical-Sum is presented. The dataset was created using MTS-Dialogues and ACI-Bench Translations for clinical summarization research. The dataset was created based on western data. That's why it was not generalize well to Bangladesh. As far we studied, very limited literature is devoted to medical query summarization. There are just two datasets available for the task:

'MeQSum' Ben Abacha and Demner-Fushman [3] and 'CHQ-Summ' Yadav, Gupta, and Demner-Fushman [19], which solely contain English text. The absence of work on medical query summarization is a significant constraint for the domain, as domain-specific models are known to outperform general ones Trewartha, Walker, Huo, *et al.* [18].

To summarise consumer health issues, the initial dataset was MeQSum, which included 1,000 samples obtained from the US National Library of Medicine. There was just one dataset available for summarising medical questions at the time, therefore it was restricted. In an effort to tackle the scarcity of datasets, the 'CHQ-Summ' dataset was introduced by Yadav, Gupta, and Demner-Fushman [19]. A total of 1,507 samples were extracted from the Yahoo community forum's question-answering section for this dataset. The variety of queries that medical practitioners may encounter is better reflected in informal data sources, which also promote diversity.

The lack of linguistic diversity is a notable drawback of existing literature. Medical query summary could greatly benefit overcrowded areas like Bangladesh, where health-care workers are sometimes overburdened by patient volume Razu, Yasmin, Arif, *et al.* [14]. Khan, Kamal, Chowdhury, *et al.* [10] contributes towards solving this issue by presenting the first CHQ summarization dataset, 'BanglaCHQSumm', for the Bangla language. This is the base literature of our work.

In this study Khan, Kamal, Chowdhury, *et al.* [10], questions from a public medical forum spanning 32 health categories were collected and subsequently processed by removing sensitive information and cleaning the data using regular expressions and Named Entity Recognition (NER). Five annotators were engaged to create concise summaries from the cleaned questions. The performance of various models, including BanglaT5, mT5, and mBART, was evaluated using ROUGE and BERTScore metrics. To address the shortcomings of automated evaluation metrics, human ratings additionally assessed the generated summaries based on informativeness, conciseness, and fluency, ensuring a comprehensive evaluation of the models' effectiveness in summarizing medical queries in Bengali. The modest size of the BanglaCHQ-Summ dataset limits its robustness despite surpassing

similar English datasets.

Now speaking about the performance evaluation of the models in natural language generation tasks like machine translation, abstractive summarization, question answering etc., specially zero shot evaluation, this provides us the insight of best performing model in these specific tasks along with how well do the models generalize to unseen or complex concepts not explicitly present in the training data.

Already many works have been conducted on the performance evaluation of the SOTA fine tuned models on tasks like machine translation, abstractive summarization, question answering Işıkdemir [8]. Few works on performance evaluation of models in the previous mentioned tasks for low resource language like bangla is done already. Mostly the performance evaluation is done of the bangla natural language generation model BanglaT5 and multilingual model like mT5 using metrics like ROUGE, BLEU Score, BERT Score etc. Adib and Alam [1], Bhattacharjee, Hasan, Ahmad, *et al.* [4], and Khan, Kamal, Chowdhury, *et al.* [10].

A thorough evaluation of NLG models in Bangla is carried out in Bhattacharjee, Hasan, Ahmad, *et al.* [4]. For the goal of pretraining, this study amassed a massive 27.5 GB Bangla text corpus. Machine translation, abstractive text summarization, and question answering were among the many tasks carried out on this massive dataset using the sequence-to-sequence Transformer known as the BanglaT5 model. Using detokenized SacreBLEU and ROUGE-2 metrics, the BanglaT5 model was thoroughly tested for its performance on several Bengali natural language processing tasks.

Though LLMs have been tested in a variety of tasks, the majority of which have been conducted in English, they have yet to be thoroughly examined in under-resourced languages such as Bengali. To this purpose, Kabir, Islam, Laskar, *et al.* [9] introduces BenLLM-Eval, which is a complete evaluation of LLMs to benchmark their performance in the Bengali language with few resources. In this regard, they chose a variety of important and diverse Bengali NLP tasks, including text summarization, question answering, paraphrasing, natural language inference, transliteration, text classification, and sentiment analysis, for zero-shot evaluation of popular LLMs, including GPT-3.5, LLaMA-2-13b-chat, and Claude-2. Their experimental results reveal that while in some Bengali NLP tasks, zero-shot LLMs could attain performance on par, if not better than existing SOTA fine-tuned models; in most tasks, their performance is rather poor.

In this regard, our work of running performance evaluation on SOTA fine tuned for bangla text summarization task models like BanglaT5 and mT5 as well as large language models like GPT-3.5 and GPT-4 on bangla medical query summarization task will guide us to

choose the best performing model in bangla medical query summarization task. Besides, it will give us comparative analysis on how the most used llm perform against the fine tuned model. Lastly, the zeroshot evaluation will give us insight on how well do the models generalize to unseen medical subdomains or complex medical concepts not explicitly present in the training data.

Recent advancements in natural language processing (NLP) have seen the development of sophisticated large language models (LLMs) like GPT-4 and BERT, which excel in various text-related tasks. Studies by researchers such as Lee et al. (2020) and Alsentzer et al. (2019) have demonstrated the potential of these models in medical applications, including clinical text mining and electronic health record (EHR) summarization. However, the majority of these studies focus on high-resource languages and often require extensive fine-tuning with domain-specific data, which is not always available for low-resource languages like Bangla.

While significant progress has been made, there are notable limitations in current research. Most existing studies emphasize fine-tuning LLMs on large, labeled datasets, which are scarce for low-resource languages. Additionally, the zero-shot capabilities of LLMs-where models are applied to new tasks without task-specific training are under-explored in the context of medical query summarization for the Bangla language. This research seeks to address these gaps by benchmarking the performance of LLMs using the 'BanglaCHQ-Summ' dataset, thereby providing new insights and methodologies for improving medical information accessibility in low-resource language settings.

Chapter 3

Proposed Methodology

This section outlines the methodological framework employed to evaluate the performance of various multilingual LLMs in addressing medical queries formulated in Bengali. The evaluation adopted a zero-shot learning paradigm, leveraging the inherent capabilities of the pre-trained models without domain-specific fine-tuning.

3.1 Dataset Selection

The selection of an appropriate dataset is paramount as it directly influences the robustness and generalizability of machine learning models. A comprehensive dataset encompassing real-world medical inquiries phrased in natural language was assembled. This dataset aimed to represent the diversity of medical questions encountered by Bengali speakers.

The BanglaCHQ-Summ dataset [10], published in 2023 is a specialized resource tailored for the task of summarizing Bangla medical queries. This dataset comprises 2,350 distinct question-summary pairs, facilitating a comprehensive analysis of medical queries in the Bangla language. Each entry within the dataset includes a question and its corresponding summary, providing valuable pairs for training and evaluation purposes. The average length of the original questions stands at 326 words, indicating the complexity and detail typical of medical inquiries. In contrast, the summaries average 136 words, reflecting a significant reduction in length while retaining essential information, thus demonstrating the summarization capability. Each sample in the dataset is systematically annotated with identifiers (ID), the original question, the generated summary, and relevant indices, enabling structured data manipulation and retrieval. This dataset serves as a critical tool for developing and testing summarization models in the context of Bangla medical language, contributing to advancements in natural language processing (NLP) and healthcare

information systems.

3.2 Model Selection and Zero-Shot Learning Approach

To evaluate zero-shot capabilities on Bangla medical health queries, four prominent models were selected: BanglaT5, mT5, GPT-3.5, and GPT-4. These models represent a spectrum of advanced natural language processing technologies with varying design principles and capabilities, tailored to handle complex linguistic tasks across different domains and languages.

BanglaT5, introduced by csebuetnlp in 2017 for Bengali language tasks, utilizes the state-of-the-art T5 transformer architecture, ensuring effective management of Bangla's nuances and complexities. Its architecture excels in capturing context and generating coherent text, making it suitable for the evaluation task.

mT5, a multilingual extension of T5, offers insights into how a model trained across multiple languages performs on Bangla-specific queries, despite lacking direct Bangla-centric training. This comparison aids in understanding multilingual model adaptability and robustness in language-specific tasks.

GPT-3.5, an autoregressive language model by OpenAI with 175 billion parameters, serves as a robust baseline for generative models in zero-shot evaluations on Bangla queries. Its architecture excels in generating contextually relevant text, highlighting its efficacy in specialized linguistic contexts.

GPT-4, a multimodal large language model with 1.76 trillion parameters, represents the forefront of AI capabilities, offering unparalleled processing and text generation across languages and tasks. Including GPT-4 explores state-of-the-art model capabilities in handling Bangla medical health queries without prior fine-tuning, shedding light on their generalization capabilities and limitations in specialized linguistic contexts.

This selection of models, encompassing both language-specific and general-purpose architectures, facilitates a comprehensive evaluation of zero-shot performance on Bangla medical health queries, emphasizing strengths and areas for improvement across different modeling approaches.

3.3 Benchmarking and Evaluation Framework

The Benchmarking and Evaluation Framework serves as a crucial yardstick for assessing the performance and capabilities of Language Model Models (LLMs), ensuring their effectiveness in real-world applications.

3.3.1 Inference Pipeline for LLM Evaluation

An inference pipeline was established for each LLM, encompassing tokenization, embedding, decoder processing, and post-processing of model outputs.

- **Tokenization:** As mentioned earlier, the Bengali medical queries were likely tokenized into subword units suitable for the respective LLM's architecture. This step ensures compatibility with the model's vocabulary and internal representations.
- **Embedding:** The tokenized queries were then mapped to dense vector representations using the model's pre-trained embedding layer. This process allows the model to capture semantic relationships within the query and translate it into a format suitable for further processing.
- **Decoder Processing:** The embedded representations were fed into the decoder portion of the LLM architecture. Through a series of attention mechanisms and activation functions, the decoder would attempt to generate a response sequence, conditioned on the prompt and the encoded query representation. This stage essentially involves the model leveraging its pre-trained knowledge to generate a response that aligns with the prompt and the encoded query.

3.3.2 Post-Processing of Model Outputs:

The raw model outputs have undergone post-processing steps to refine the generated responses and improve their fluency and coherence. Techniques like top-k sampling (selecting the k most probable tokens at each step) have been employed.

3.3.3 Evaluation Metrics for Response Quality

The quality of the generated responses was assessed using automatic evaluation metrics such as ROUGE-Score, BERTscore, and BLEU.

ROUGE-Score

To compare the accuracy of machine-generated text with that of reference replies, a set of metrics called ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is often utilised. It comes in several versions:

- **ROUGE-N (N-gram overlap):** Measures overlap of n-grams (sequences of n con-

tiguous words) between the generated response G and the reference answers R .

$$\text{ROUGE-N}_n(G, R) = \frac{\sum_{g \in G} \min(\text{count}_{n\text{-gram}}(g, G), \text{count}_{n\text{-gram}}(g, R))}{\sum_{g \in G} \text{count}_{n\text{-gram}}(g, G)}$$

- **ROUGE-L (Longest Common Subsequence):** Finds the subsequence that has the longest common ancestor between G and R , considering word tokens.

$$\text{ROUGE-L}(G, R) = \frac{\text{LCS}(G, R)}{|R|}$$

- **ROUGE-W (Weighted ROUGE):** Accounts for the importance of words through weighting schemes.

ROUGE scores typically range from 0 to 1, where higher scores indicate better similarity between the generated response and the reference answers.

BERTscore

BERTscore evaluates the quality of generated responses by measuring the similarity of token embeddings from BERT (Bidirectional Encoder Representations from Transformers) between the generated response G and the reference answers R .

$$\text{BERTscore}(G, R) = F_1\text{-score of } (\{\text{BERT}_{\text{embed}}(g_i)\}, \{\text{BERT}_{\text{embed}}(r_j)\})$$

Here, $\text{BERT}_{\text{embed}}(x)$ denotes the BERT embeddings of tokens in x , and F_1 -score measures the harmonic mean of precision and recall.

BLEU (Bilingual Evaluation Understudy)

BLEU uses n-gram overlap to determine how similar the produced response G is to the reference replies R . It determines the accuracy of n-grams up to a given order, usually four grammes. between G and R .

$$\text{BLEU}(G, R) = \text{BP} \times \exp\left(\sum_{n=1}^N \omega_n \log p_n\right)$$

Where BP (Brevity Penalty) penalizes shorter responses, p_n is the precision of n-grams, and ω_n are weights typically set to $\frac{1}{N}$ where N is the maximum n-gram order considered.

BLEU score ranges from 0 to 1, with higher scores indicating better similarity between G and R .

These metrics collectively provide insights into the quality of machine-generated responses in terms of factual accuracy, semantic similarity, and linguistic overlap with respect to reference answers.

3.4 Analysis and Interpretation of Results

The performance of each LLM across different evaluation metrics was comparatively analyzed to identify strengths and weaknesses.

Factual Accuracy: Models with high ROUGE-L scores might demonstrate a stronger ability to provide factually accurate responses.

Semantic Similarity: High BERTScore could indicate that the generated responses effectively capture the intended meaning of the queries.

Fluency: Models with high BLEU scores might be more adept at producing fluent and grammatically correct responses.

Chapter 4

Results and Discussion

The first example reveals that GPT-4, GPT3.5, and BanglaT5 incorporate the supplementary medical data, an essential piece of information, according to Table ???. Unfortunately, the mT5 was unable to record all relevant details about the query in this particular sample. However, the second sample reveals that the GPT-3.5 big language model committed a significant mistake. The patient reported abdominal discomfort, which was summarised as leg cramp by GPT-3.5 but not by BanglaT5 or GPT-4. Even here, mT5 failed to adequately summarise the query. While mT5, the bilingual model, eliminates repetitious parts from their summaries, BanglaT5 and GPT-4 keep them in while properly reflecting the core complaint. As a result, the GPT-3.5 occasionally provides summary results that are not useful. This is what we saw when we looked at the model's summary output. We find that GPT-4 and BanglaT5 provide superior summaries when compared. Since they do better on the zero shot evaluation metrics as well, this finding has not altered.

Table 4.1: Samples of summaries generated by the benchmark models

Sample 1	
Original Question	আমার ছেলে বয়স 25 মাস । তাঁর খিচুনি আসে পতি সাপতায় । এখনও বসতে পারে না , কথা বলতে পারে না , ঘর সোজা হয়ে বেশি খন রাখতে পারে না ঔষধ চলে , বারবিট 30 , ইরাসেট 250 , কিন্তু খিচুনি বন্ধ হয় না । দয়া করে আমাকে ভাল একটা পরামর্শ দিবেন ।
Annotated Summary	২৫ মাস বয়সী ছেলে , প্রতি সপ্তাহে খিচুনি হয় । কথা বলতে , ঘাড় সোজা রাখতে , বসতে পারেন । বারবিট ও ইরাসেট সেবন করছে কিন্তু খিচুনি বন্ধ হচ্ছেনা । কি করণীয় ?
GPT-4 Summary	২৫ মাস বয়সী একটি ছেলের প্রতি সপ্তাহে খিচুনি হয়, বসতে পারে না, কথা বলতে পারে না এবং ঘাড় সোজা রাখতে পারে না । বারবিট ৩০ এবং ইরাসেট ২৫০ ওষুধ চললেও খিচুনি বন্ধ হচ্ছে না । ভালো পরামর্শ চাওয়া হয়েছে ।
GPT-3.5 Summary	ছেলের বয়স ২৫ মাস । সাধারণভাবে পায়খানা অসুস্থতার লক্ষণ আসে । তিনি এখনও বসতে পারছেন না, কথা বলতে পারছেন না, পা সোজা রাখতে পারছেন না । ঔষধ হিসেবে বারবিট ৩০ এবং ইরাসেট ২৫০ ব্যবহার করা হচ্ছে, তবে পায়খানি বন্ধ হয়নি । ভালো পরামর্শ চাই ।
BanglaT5 Summary	আমার ২৫ মাসের একটি ছেলে সাপ্তাহিক খিচুনি, বসতে না, কথা বলতে না, এবং ঘাড় সোজা রাখতে পারে না । বারবিট ৩০ এবং ইরাসেট ২৫০ ঔষধ চললেও খিচুনি বন্ধ হচ্ছে না ।
mT5 Summary	আমার ছেলে খিচুনি, কিন্তু এখনও বসতে পারে না । তাঁর ঔষধ চলে না ।
Sample 2	
Original Question	আমার বয়স ২৫ বছর । আজ কিছুদিন যাবত আমার ঘন ঘন টয়লেট হচ্ছে । হটাৎ পেটের মধ্যে ব্যাথা অনুভূত হয়ে টয়লেটে চাপ দেয় এবং টয়লেটে যাওয়ার পরে ও টয়লেট ক্লিয়ারলি হচ্ছে না । এই সমস্যাটার সমাধান জানাবেন । আর একটি সমস্যা আছে , সেটি হল আজ দুই দিন যাবত হটাৎ করে আমার ঠান্ডা লেগে সর্দি + অনেক হাঁচি হচ্ছে ।
Annotated Summary	বয়স ২৫ । ঘন ঘন টয়লেট হচ্ছে । পেটের মধ্যে ব্যাথা , টয়লেট ক্লিয়ারলি হচ্ছে না । আজ দুই দিন যাবত ঠান্ডা লেগে সর্দি , হাঁচি হচ্ছে ।
GPT-4 Summary	২৫ বছর বয়সী একজন ব্যক্তির ঘন ঘন পায়খানা হচ্ছে এবং পায়খানার পরেও পুরোপুরি পরিষ্কার হচ্ছে না । সাম্প্রতিককালে ঠান্ডা, সর্দি এবং হাঁচির সমস্যাও দেখা দিয়েছে ।
GPT-3.5 Summary	আমার একটি পাঁচটি সমস্যা আছে: টয়লেট সমস্যা, হাতে রগে টান খায়, ঠান্ডা লেগে সর্দি ও অনেক হাঁচি ।
BanglaT5 Summary	আমি ২৫ বছরের । আজ টয়লেটে যাওয়ার পর দেখেছি আমি ঘন পায়খানা হচ্ছে এবং পায়খানার পরেও পুরোপুরি পরিষ্কার অনুভব করছি না । সাম্প্রতিককালে আমার ঠান্ডা, সর্দি এবং হাঁচির সমস্যাও দেখা দিয়েছে । এই সমস্যাগুলোর সমাধান সম্পর্কে কিছু জানালে খুব উপকার হবে ।
mT5 Summary	আমার বয়স ২৫ বছর । আমি আজ টয়লেটে যাওয়ার পর এই সমস্যার সমাধান জানতে চাইছি, কিন্তু এটা নিয়ে কী কিছু জানেন?

The results of the zero-shot evaluation of different state-of-the-art models on the BanglaCHQ-Summ dataset are shown in tables 4.2 and 4.3. These tables concentrate on the models' abilities to summarise Bengali medical queries. The metrics score, which includes R1, R2, RL, BS, and BLEU, gives us one view of the models' output quality, but each measure offers a unique viewpoint.

4.1 Analysis and Comparison

1. ROUGE-1 (R1):

- GPT-4 achieves the highest R1 score (29.22), significantly outperforming the other models.
- BanglaT5 (18.28) and mT5 (17.58) have comparable scores, with BanglaT5 slightly ahead.
- GPT-3.5 (15.16) scores the lowest among the four models in this metric.

2. ROUGE-2 (R2):

- Again, GPT-4 leads with a score of 18.80, indicating its superior capability in capturing the sequence of words.
- BanglaT5 (11.15) and mT5 (10.65) follow, with similar performances.
- GPT-3.5 scores the lowest (8.98), reflecting challenges in bigram generation.

3. ROUGE-L (RL):

- GPT-4 maintains its lead with 27.71, showing its effectiveness in generating coherent and relevant sequences.
- BanglaT5 (16.33) and mT5 (15.60) have close scores, with BanglaT5 slightly outperforming mT5.
- GPT-3.5 (14.05) again has the lowest score, indicating less alignment with the reference sequences.

4. BERTScore (BS):

- GPT-4 excels with a BERTScore of 90.25, highlighting its ability to generate semantically accurate summaries.
- GPT-3.5 (86.95), BanglaT5 (86.09), and mT5 (85.85) have relatively close scores, but GPT-3.5 slightly outperforms the other two.

5. BLEU Score:

- GPT-4 achieves the highest BLEU score of 0.086.
- GPT-3.5, BanglaT5 and mT5 scored considerably lower than GPT-4 with a BLEU scores of 0.031, 0.039 and 0.035 respectively.

Table 4.2: Automatic evaluation results of BanglaCHQ-Summ

Model	R1	R2	RL	BS
GPT-4	29.22	18.80	27.71	90.25
GPT-3.5	15.16	8.98	14.05	86.95
BanglaT5	18.28	11.15	16.33	86.09
mT5	17.58	10.65	15.60	85.85

Table 4.3: BLEU Score evaluation results of BanglaCHQ-Summ

Model	BLEU Score
GPT-4	0.086
GPT-3.5	0.031
BanglaT5	0.039
mT5	0.035

If we summarize the result we find, GPT-4 consistently outperforms all other models across all metrics, demonstrating its superior performance in summarizing Bengali medical queries. Its high scores across ROUGE, BLEU and BERTScore metrics indicate both syntactic and semantic strengths. BanglaT5 and mT5 show similar performance levels, with BanglaT5 having a slight edge in R1 and RL metrics, while mT5 follows closely. GPT-3.5 generally scores the lowest across all metrics, though it still provides useful insights into its capabilities and areas for improvement.

Chapter 5

Conclusion

In this chapter, we summarize our comparative analysis of large language models and traditional machine learning models, and offer directions for future research. Our findings reveal both the strengths and limitations of these models in handling Bangla medical health queries, laying a foundation for future advancements. We also explore potential avenues for enhancing model robustness, integrating multimodal data, and expanding evaluation frameworks, thereby setting the stage for continued progress in this domain.

5.1 Conclusion

To further understand the efficacy and possible uses of Bengali language models, this study investigated their zero-shot assessment for the job of medical query summarization. Our investigation into the efficacy of cutting-edge models in summarising Bengali medical concerns included extensive testing and analysis of BanglaT5, mT5, GPT-3.5, and GPT-4. Our research showed that GPT-4 performed the best, with a BERTScore of 90.25%, proving that it effectively generates brief and accurate summaries. These findings demonstrate the great promise of sophisticated language models for dealing with the difficulties of medical query summarising tasks, especially in languages like Bengali that have limited resources.

Our research not only shows how effective Bengali language models are, but it also explains how this has wider ramifications for healthcare providers and their patients. Healthcare consultations, patient-provider communication, and health outcomes for Bengali-speaking populations could all be improved with the help of efficient summarization tools that streamline the process of extracting essential information from medical queries. Academic and practical fields alike can benefit from the evaluation's wealth of information

for future studies on Bengali medical inquiry summarising.

5.2 Future Work

Despite the valuable insights gained, it is essential to acknowledge the limitations of this study and outline potential directions for future exploration. Several avenues for research and development emerge:

5.2.1 Enhancement of Model Robustness

Future research should focus on enhancing the robustness of models, particularly in handling the nuances and intricacies specific to Bangla Consumer Health Questions (CHQs). This can be achieved through further fine-tuning on domain-specific data or exploring transfer learning techniques that adapt models to Bangla language peculiarities.

5.2.2 Integration of Multimodal Approaches

Integrating multimodal approaches that incorporate both textual and visual information could further enhance the accuracy and relevance of summarization outputs. By leveraging additional contextual cues from images or diagrams related to medical queries, models can provide more informative and contextually rich summaries.

5.2.3 Exploration of Zero-shot Learning Strategies

Further exploration of zero-shot learning strategies is warranted to improve the adaptability of models to unseen medical queries and diverse linguistic variations within the Bangla-speaking community. Investigating novel methodologies that minimize the need for extensive fine-tuning while maximizing performance on diverse datasets could significantly broaden the applicability of automated summarization systems.

5.2.4 Evaluation on Larger and Diverse Datasets

Scaling up evaluations to larger and more diverse datasets beyond 'BanglaCHQ-Summ' will be crucial to validate the generalizability and robustness of the evaluated models. This includes datasets spanning various medical specialties, different dialects of Bangla, and a broader spectrum of medical query types to capture real-world variability.

5.2.5 User-Centric Evaluation and Deployment

Conducting user-centric studies to assess the practical utility and user acceptance of automated summarization systems in real-world healthcare settings is essential. Evaluating how these systems impact healthcare professionals' workflow efficiency and patient satisfaction can provide valuable insights for refining and optimizing system deployment strategies.

5.2.6 Ethical and Societal Implications

Exploring the ethical implications of deploying automated summarization systems in healthcare contexts is imperative. Future research should address concerns related to privacy, data security, bias mitigation, and ensuring equitable access to healthcare information across diverse socio-economic backgrounds within the Bangla-speaking population.

Finally, by shedding light on the potential of Bengali language models for medical query summarising, our research helps move the natural language processing and healthcare industries forward. Improving the availability and quality of healthcare information and services is the ultimate goal of addressing the stated future research areas, which will aid in the continuous development of digital healthcare solutions adapted to the unique requirements of Bangla-speaking populations. Moving forward, it is critical to investigate novel ways to use NLP technology to enhance healthcare accessibility, efficiency, and quality in settings with a high concentration of language minorities.

References

- [1] Q. A. R. Adib and S. B. Alam, “Bnclinical-sum: Benchmarking datasets for bangla long & short clinical dialogue summarization,” Ph.D. dissertation, Brac University, 2024.
- [2] S. Akter, A. S. Asa, M. P. Uddin, M. D. Hossain, S. K. Roy, and M. I. Afjal, “An extractive text summarization technique for bengali document (s) using k-means clustering algorithm,” in *2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (ICIVPR)*, IEEE, 2017, pp. 1–6.
- [3] A. Ben Abacha and D. Demner-Fushman, “On the summarization of consumer health questions,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2228–2234. DOI: 10.18653/v1/P19-1215. [Online]. Available: <https://aclanthology.org/P19-1215>.
- [4] A. Bhattacharjee, T. Hasan, W. U. Ahmad, and R. Shahriyar, “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*, A. Vlachos and I. Augenstein, Eds., Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 726–735. DOI: 10.18653/v1/2023.findings-eacl.54. [Online]. Available: <https://aclanthology.org/2023.findings-eacl.54>.
- [5] M. Gambhir and V. Gupta, “Recent automatic text summarization techniques: A survey,” *Artificial Intelligence Review*, vol. 47, pp. 1–66, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:24465182>.
- [6] M. M. Haque, S. Pervin, A. Hossain, and Z. Begum, “Approaches and trends of automatic bangla text summarization: Challenges and opportunities,” *IJTD*, vol. 11, no. 4, pp. 1–17, 2020. DOI: 10.4018/IJTD.20201001.oa. [Online]. Available: <https://doi.org/10.4018/IJTD.20201001.oa>.
- [7] S. M. A. I. Hayat, A. Das, and M. M. Hoque, “Abstractive bengali text summarization using transformer-based learning,” in *2023 6th International Conference*

- on *Electrical Information and Communication Technology (EICT)*, 2023, pp. 1–6. DOI: 10.1109/EICT61409.2023.10427906.
- [8] Y. E. Işıkdemir, “Nlp transformers: Analysis of llms and traditional approaches for enhanced text summarization,” *Eskişehir Osmangazi Üniversitesi Mühendislik ve Mimarlık Fakültesi Dergisi*, vol. 32, no. 1, pp. 1140–1151, 2024.
- [9] M. Kabir, M. S. Islam, M. T. R. Laskar, M. T. Nayeem, M. S. Bari, and E. Hoque, “Benllmeval: A comprehensive evaluation into the potentials and pitfalls of large language models on bengali nlp,” *arXiv preprint arXiv:2309.13173*, 2023.
- [10] A. Khan, F. Kamal, M. A. Chowdhury, T. Ahmed, M. T. R. Laskar, and S. Ahmed, “BanglaCHQ-summ: An abstractive summarization dataset for medical queries in Bangla conversational speech,” in *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, F. Alam, S. Kar, S. A. Chowdhury, F. Sadeque, and R. Amin, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 85–93. DOI: 10.18653/v1/2023.banglalp-1.10. [Online]. Available: <https://aclanthology.org/2023.banglalp-1.10>.
- [11] Y. Kumar, K. A. Kaur, and S. Kaur, “Study of automatic text summarization approaches in different languages,” *Artificial Intelligence Review*, vol. 54, pp. 5897–5929, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:233936657>.
- [12] M. Moradi and N. Ghadiri, “Text summarization in the biomedical domain,” *ArXiv*, vol. abs/1908.02285, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:199472647>.
- [13] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, “A survey of automatic text summarization: Progress, process and challenges,” *IEEE Access*, vol. 9, pp. 156 043–156 070, 2021. DOI: 10.1109/ACCESS.2021.3129786.
- [14] S. R. Razu, T. Yasmin, T. B. Arif, *et al.*, “Challenges faced by healthcare professionals during the covid-19 pandemic: A qualitative inquiry from bangladesh,” *Frontiers in public health*, vol. 9, p. 647 315, 2021.
- [15] A. Sarkar and M. S. Hossen, “Automatic bangla text summarization using term frequency and semantic similarity approach,” in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, 2018, pp. 1–6. DOI: 10.1109/ICCITECHN.2018.8631934.
- [16] A. Sarker, Y.-C. Yang, M. A. Al-garadi, and A. Abbas, “A light-weight text summarization system for fast access to medical evidence,” *Frontiers in Digital Health*, vol. 2, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:227257712>.

- [17] M. A. I. Talukder, S. Abujar, A. K. M. Masum, F. Faisal, and S. A. Hossain, “Bengali abstractive text summarization using sequence to sequence rnns,” in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019, pp. 1–5. DOI: 10.1109/ICCCNT45670.2019.8944839.
- [18] A. Trewartha, N. Walker, H. Huo, *et al.*, “Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science,” *Patterns*, vol. 3, p. 100488, Apr. 2022. DOI: 10.1016/j.patter.2022.100488.
- [19] S. Yadav, D. Gupta, and D. Demner-Fushman, *Chq-summ: A dataset for consumer healthcare question summarization*, Jun. 2022. DOI: 10.48550/arXiv.2206.06581.