BanglaCHQ-Summ: An Abstractive Summarization Dataset for Medical Queries in Bangla Conversational Speech

Alvi Aveen Khan[1], **Fida Kamal**[1], **Md. Abrar Chowdhury**[1],
Tasnim Ahmed[1,2], **Md. Tahmid Rahman Laskar**[3], and **Sabbir Ahmed**[1]
[1]Islamic University of Technology, [2]Queen's University, [3]York University
[1]{alviaveen, fidakamal, abrar35, sabbirahmed}@iut-dhaka.edu
[2]tasnim.ahmed@queensu.ca, [3]tahmid20@yorku.ca

Abstract

Online health consultation is steadily gaining popularity as a platform for patients to discuss their medical health inquiries, known as Consumer Health Questions (CHQs). The emergence of the COVID-19 pandemic has also led to a surge in the use of such platforms, creating a significant burden for the limited number of healthcare professionals attempting to respond to the influx of questions. Abstractive text summarization is a promising solution to this challenge, since shortening CHQs to only the information essential to answering them reduces the amount of time spent parsing unnecessary information. The summarization process can also serve as an intermediate step towards the eventual development of an automated medical question-answering system. This paper presents 'BanglaCHQ-Summ', the first CHQ summarization dataset for the Bangla language, consisting of 2,350 question-summary pairs. It is benchmarked on state-of-the-art Bangla and multilingual text generation models, with the best-performing model, BanglaT5, achieving a ROUGE-L score of 48.35%. In addition, we address the limitations of existing automatic metrics for summarization by conducting a human evaluation. The dataset and all relevant code used in this work have been made publicly available[1].

1 Introduction

The answers to general health inquiries can often be obtained by utilizing internet search engines, but addressing queries by individual users in a manner that caters to their specific circumstances remains a manual process. Such queries, known as Consumer Health Questions (CHQs), are frequently found on online health forums, and answering them is becoming increasingly time-consuming and labour-intensive for medical professionals (Ma et al., 2018). The task is made even more difficult by the fact that patients are often overly descriptive when asking questions, providing unnecessary details (Roberts and Demner-Fushman, 2016). The ability to identify and discard these unnecessary details would save a lot of time for the response providers and would also be an important step towards the eventual development of an automated question-answering system (Abacha and Demner-Fushman, 2019a).

Abstractive text summarization is the task of generating a shortened and human-readable version of the original text that retains the important information (Nallapati et al., 2016). Despite the recent improvement in this domain due to the development of transformer-based architectures as well as the greater availability of data, progress has been somewhat limited in CHQ summarization (Abacha and Demner-Fushman, 2019b; Yadav et al., 2022a, 2021). This shortcoming is particularly notable for low-resource languages like Bangla (Alam et al., 2021), for which there is no existing work on this task.

Developing a dataset dedicated to this language presents a substantial challenge to existing architectures for several reasons. Firstly, Bangla is an exceedingly complicated language in comparison to English, allowing for more flexible sentence structuring (Sinha et al., 2016) and a significantly greater number of inflections (220 as opposed to just 9 in English (Bhattacharya et al., 2005)), resulting in noisier text. Furthermore, the diversified dialects exacerbate the issue, with the language used in one region frequently being entirely unintelligible in another (Shahed, 1993). Navigating this complexity and successfully identifying the relevant medical information is a significantly complicated task.

Unfortunately, the Bangla text summarization architectures currently available do not account for the complications of informal speech in medical contexts since they were mostly trained on

[1]https://github.com/alvi-khan/BanglaCHQ-Summ

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 85–93

news articles (Bhattacharjee et al., 2023; Hasan et al., 2021), making them unsuitable for summarizing medical text. This paper addresses the lack of medically relevant data by introducing the first human-annotated Bangla CHQ summarization dataset, 'BanglaCHQ-Summ', consisting of 2350 question-summary pairs. The data was collected from a public online health forum used by hundreds of native Bangla speakers, allowing it to present an accurate representation of the health questions that are generally present on online forums. In addition to the dataset, we also discuss the shortcomings of established evaluation metrics of text summarization tasks and explore a methodology for human evaluation that addresses the shortcomings.

2 Related Work

Although a large amount of work has been dedicated to text summarization in general (Allahyari et al., 2017; Nenkova and McKeown, 2012; El-Kassas et al., 2021), very limited literature is devoted to CHQ summarization. To the best of our knowledge, there are only two datasets available for the task, 'MeQSum' (Abacha and Demner-Fushman, 2019b) and 'CHQ-Summ' (Yadav et al., 2022b), and both consist exclusively of English text. The lack of work addressing CHQ summarization is a major limitation for the domain since domain-specific models are known to outperform general ones (Trewartha et al., 2022).

MeQSum was the first dataset for consumer health question summarization, consisting of 1,000 samples collected from the U.S. National Library of Medicine. The dataset has a relatively small size but was also the only medical question summarization dataset available at that time. Yadav et al. (2022b) attempted to address the lack of available datasets by introducing the 'CHQ-Summ' dataset. This dataset consists of 1,507 samples collected from the Yahoo community question-answering forum. The informal source of the data enhances its diversity and presents a more realistic depiction of the questions that medical professionals are likely to encounter.

A notable shortcoming of the existing literature is the lack of diversity in language. The advantages of CHQ summarization should prove extremely beneficial if its application can be extended to support overpopulated regions such as Bangladesh, where healthcare workers are frequently overwhelmed by the volume of patients (Razu et al., 2021). Introducing a Bangla dataset contributes towards solving this issue, and the knowledge gained is also transferable to other Indo-Aryan languages of the Indian subcontinent.

3 The BanglaCHQ-Summ Dataset

In this section, we demonstrate how we curated the proposed BanglaCHQ-Summ dataset.

3.1 Data Collection

We collected the questions from a renowned medical forum[2] that publicly releases questions posted by users, along with answers provided by medical professionals. Given that the data was collected from a public health forum, it can be reasonably assumed that the user base consisted of individuals with average medical knowledge. This user base consists of individuals from diverse backgrounds based on the linguistic variety of the questions, which is a particularly strong point for the dataset since it presents an accurate representation of the variety of the Bangla language discussed earlier. The forum contains questions belonging to a total of 32 categories, which allows the samples to cover a broad spectrum of health issues. However, the information related to the categories has been omitted from our dataset as the category assignment is done by the patients while posting the queries, which can often be inaccurate.

3.2 Pre-Processing

A portion of the collected data contained sensitive information. To protect the privacy of the patients, such personally identifiable information has been removed by utilizing regular expressions to identify email addresses and phone numbers and a Bangla Named Entity Recognition model[3] to identify names. The data was then also inspected manually. Additionally, duplicate entries, URLs, and spam text were also removed as part of the overall data-cleaning process.

3.3 Annotation

A team of 5 annotators with at least an undergraduate level of education was chosen after carefully evaluating their summarization capabilities in the Bangla language. The primary instruction provided to the annotators was to make the text as

[2] https://daktarbhai.com/

[3] https://pypi.org/project/bnlp-toolkit/

Question	I have chronic stress and anxiety, I am loosing everything in my life, but do not want pills, what can I do? I have problems with stress, however it is not just that, but the fact that every time I start with this condition it turns into a huge fear of choking and my mind starts telling me not to eat. The last time it happened I did not eat anything solid for four months and I suffered severe damage in other parts of my body like my stomach and my heart which is worst. This time it is starting again and I am two weeks under this condition. The last time I was using antidepressants and other drugs, but when I tried cutting them the anxiety made me feel worst. This is why I changed my treatment, now I use relaxation exercises with the help on my doctor. The last time it helped me a lot, but this time I think I need more help. I am taking meditation and tai chi courses and I am expecting to take yoga classes as well. The problem is that this is taking away my life, I have doubts on whether I will be cured one day or if it will take so long that everything I have now will be lost. I need help.
Summary	What are possible non-drug treatments for chronic stress and anxiety?

Table 1: Sample summary from the CHQ-Summ dataset (Yadav et al., 2022b)

concise as possible without discarding any information essential to answer the question accurately. The complete set of guidelines is provided in Appendix A.1.

Appendix A.2 showcases a few samples of the annotated summaries from the dataset. Each annotator was provided with a set of 500 questions, among which 6% was common. The summaries of the common questions were later used to calculate the inter-annotator agreement (IAA) using the ROUGE-L metric (Lin, 2004), with the average score being 50.11%. However, this score does not take semantic differences into account, an issue previously highlighted by Yadav et al. (2022b) when they found a similar score for their work. A manual evaluation of the summaries clearly demonstrates significant semantic overlap. To quantify this, we refer to the BERTScore metric (Zhang et al.), which calculates the semantic similarity between sentences. The average BERTScore for the common questions provided to the annotators is 90.84%. Hence, we conclude that despite there being differences in the phrasing used by the annotators, the content of their summaries is largely the same.

A portion of the annotated summaries, specifically the portion used as the test set for evaluation of the benchmark models, was further verified by a physician, who determined whether the annotated summaries were appropriate and medically relevant. Based on this, we found that they strongly agreed with 80% of the annotated summaries, with only minor issues being found in the remaining 20%, which they assured us do not make the summaries inaccurate.

3.4 Dataset Attributes

The final dataset consists of 2350 question-summary pairs. The average length of the original questions was 326 words, compared to the average length of 136 words for the annotated summaries. This large difference in lengths provides evidence of the fact that users on health forums tend to ask overly descriptive questions, which in turn require more effort to parse.

The annotated summaries from our dataset are noticeably longer than those found in existing work. This difference is deliberate. Analyzing the MeQSum and CHQ-Summ datasets, we found that they prioritized shorter lengths over information retention. An example of this is provided in Table 1, which shows a sample summary taken directly from the CHQ-Summ dataset. Although the annotated summary addresses the main question asked, it leaves out a large number of additional details, such as the patient having past issues with stress to the extent of not eating solids and that they have tried using antidepressants and other drugs. The summary only allows for a generic response without considering the patient's specific circumstances. To avoid this, our annotators were instructed to retain all medically relevant information. This allows us to obtain shortened questions while still addressing the specific situation being faced by the patient.

An important finding of the summarization process was that a significant portion of patients explicitly mentioned being unable to visit medical professionals in-person during the COVID-19 pandemic. Analyzing the data from the online platform reveals a correlation, visualized in Fig. 1. The diagram compares the daily count of questions posed on the platform with the number of

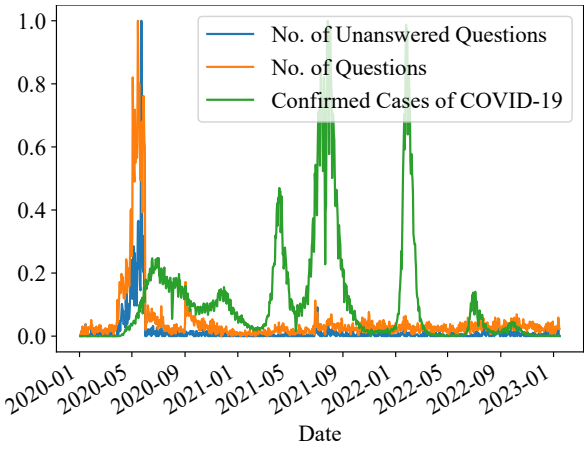

Figure 1: Comparison of time-frames for a rise in question count and unanswered questions on the online health platform with the rise in new cases of COVID-19 in Bangladesh. Values are normalized due to the large difference in scale.

confirmed cases of COVID-19 in Bangladesh[4], where the majority of the user base of the online platform resides. During the initial wave of the pandemic, there was a significant rise both in the number of questions asked and the number of questions remaining unanswered. The trend does not repeat itself during the latter waves, presumably due to the general public becoming well-informed by that time. This finding reinforces the need for Bangla CHQ summarization and, ideally, an automated question-answering system (Laskar et al., 2020) to provide support to the medical staff in unprecedented scenarios such as a pandemic.

4 Experiments

To benchmark model performance on our proposed dataset, we conduct two types of evaluation: (i) Automatic and (ii) Human. We split the dataset into training, validation, and test sets following an 80:10:10 ratio. Below, we present our findings.

4.1 Automatic Evaluation

For automatic evaluation, we experimented with one Bangla text generation model, BanglaT5 (Bhattacharjee et al., 2023), as well as two multilingual ones, mT5 (Xue et al., 2021) and mBART (Tang et al., 2020). The details of the experimental setup are provided in Appendix A.3.

To evaluate the model, we used the ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), and BERTScore (BS) metrics. For the BERTScore metric, layer 12 of the BanglaBERT (Bhattacharjee et al., 2022) model was used. The ROUGE scores measure the degree of overlap between the generated and reference summaries and are a commonly used evaluation metric for text summarization tasks. The BERTScore metric measures the semantic similarity between generated and reference summaries and is known to correlate better with human evaluation.

Model	R1	R2	RL	BS
Bangla T5	50.05	29.11	48.35	89.91
mT5	40.99	22.84	39.76	88.50
mBART	47.23	27.15	45.86	89.38

Table 2: Automatic evaluation results of BanglaCHQ-Summ

Our results, presented in Table 2, show that BanglaT5 outperforms both multilingual models on all four metrics, demonstrating that models pretrained on a language-specific corpus outperform multilingual ones.

4.2 Human Evaluation

As discussed in section 3.3, the ROUGE score gives limited insight into the quality of the generated summaries. BERTScore can better capture semantic similarities but still does not account for several important factors, such as the coherence, logical flow, or overall correctness of the text. To address these limitations, we have explored a methodology to establish quantitative metrics to evaluate summaries following Laskar et al. (2022).

To carry out this evaluation, a group of 3 annotators was provided with the same set of 30 samples along with the summaries generated by each of the three models. They rated the generated summaries on a scale of 1 to 5 based on the following metrics:

Informativeness (I): Measures the extent to which the information required to answer the question was retained in the summary. Including unnecessary information does not lower this score.

Conciseness (C): Measures how short the summary is. Including unnecessary information or being verbose while describing the necessary information lowers this score.

Fluency (F): Measures how coherent and fluent the summary is.

Table 4 shows the average score assigned to each model based on the evaluation of the three annotators. The evaluation process revealed

[4] https://covid19.who.int/region/searo/country/bd

Sample 1	
Original Question	আমার বয়স ৩০। অনেক দিন ধরে কোমরে ব্যথা করে। আমি ডাঃ দেখিয়েছি। এক্সরে করানোর পর আমাকে ন্যাপ্রক্সিন ১০ দিন দু বেলা করে খেতে দেয়। আর সাথে ক্যালসিয়াম, নিবলন খেতে দেয় ১ মাস। কিন্তু এ পর্যন্ত ব্যাথা যায় নাই। কোমরে প্রেশার দিলে ব্যাথা করে। এখন কি করা যায়।
BanglaT5 Summary	বয়স ৩০। কোমরে ব্যথা। ডাক্তার ন্যাপ্রক্সিন ১০ দিন দু বেলা করে দেয়। সাথে ক্যালসিয়াম, নিবলন খেতে দেয় ১ মাস। কিন্তু ব্যাথা যায় নাই। প্রেসার দিলে ব্যাথা করে। কি করনীয়?
mT5 Summary	বয়স ৩০। অনেক দিন ধরে কোমরে ব্যথা করে
mBART Summary	বয়স ৩০। কোমরে ব্যাথা। কি করনীয়?
Sample 2	
Original Question	আমার আম্মুর পিঠের ঠিক মাঝে ভীষণ জ্বলে প্রায় এক বসর ধরে এমন হয়। ডাক্তার দেখানো হয়েছে শুধু গ্যাস্ট্রিকের ওষুধ দেয়। কিন্তু কোন কাজ হয়না। এখন খুব জ্বলে। মাঝে মাঝে একটু কম থাকে। কোন বিশেষজ্ঞ দেখালে ভাল হয়? আর কি সমস্যা হতে পারে? বিশেষজ্ঞদের পরামর্শ চাই।
BanglaT5 Summary	পিঠের ঠিক মাঝে ভীষণ জ্বলে প্রায় এক বছর ধরে। ডাক্তার গ্যাস্ট্রিকের ওষুধ দেয়। কাজ হয়না। এখন খুব জ্বলে। মাঝে মাঝে একটু কম থাকে। কোন বিশেষজ্ঞ দেখালে ভালো হবে? আর কি সমস্যা হতে পারে?
mT5 Summary	পিঠের ঠিক মাঝে ভীষণ জ্বলে প্রায় এক বসর ধরে। ডাক্তার শুধু গ্যাস্ট্রিকের ওষুধ দেয় কিন্তু কাজ হয়না
mBART Summary	আম্মুর পিঠের ঠিক মাঝে ভীষণ জ্বলে প্রায় এক বসর ধরে এমন হয়। ডাক্তার শুধু গ্যাস্ট্রিকের ওষুধ দেয়।

Table 3: Samples of summaries generated by the benchmark models

Model	I	C	F
Bangla T5	4.09	3.83	4.27
mT5	2.94	4.18	4.39
mBART	3.34	4.00	4.12

Table 4: Human evaluation results of BanglaCHQ-Summ

that summaries with high informativeness scores tended to have relatively low conciseness scores and vice versa. This indicates that the models struggled to retain all the correct information while also being concise. Amongst the models, BanglaT5 shows significant superiority in preserving required information in its summaries but has comparatively less proficiency in conciseness compared to the multilingual models. This can be demonstrated with reference to the samples of generated summaries in Table 3.

We find from Table 3 that the first sample shows a serious error made by the multilingual models. The patient complains of waist pain, which all three models capture in their summaries, but only BanglaT5 includes the additional information regarding medicine prescribed to the patient by a doctor, a critical piece of information. On the other hand, the second sample illustrates the tendency of BanglaT5 to be excessively descriptive. The patient describes a burning sensation in their back and mentions that the medicine given by doctors does not provide relief. The latter part of the question repeats this complaint, adding no new information. The summary generated by BanglaT5 accurately reflects the main complaint but retains the repetitive portions, while the summaries generated by the multilingual models exclude the repetitive portions.

5 Conclusion

In this paper, we propose the first CHQ summarization dataset for the Bangla Language. The source of the data used in the creation of the dataset also presents an advancement towards a more accurate representation of the diversity of the language. In addition, we explore a methodology for human evaluation that addresses the limitations of existing text summarization evaluation metrics. Given the sensitive nature of the public health domain, improvements in the performance of the summarization models, alongside evaluating how Large Language Models (Jahan et al., 2023) perform in this dataset could be a good direction for future research.

Limitations

One limitation of this work is that the size of the proposed dataset is quite modest. However, even the existing English question summarization datasets have limited sizes. In this regard, our dataset, although being for the low-resourced Bangla language, surpasses the sizes of similar datasets available in English.

Another limitation of this work is that, while our dataset has been benchmarked on widely used text summarization models, the use of such models assumes the availability of significant computational resources that many organizations may not be able to afford. Although utilizing computational resources from third-party institutions will likely be able to address this issue, the sensitive nature of medical data makes sharing the data with third parties an unfavorable solution.

Ethics Statement

The Consumer Health Questions (CHQs) collected to prepare our dataset are publicly available. As of October 17, 2023, the terms and conditions of the online health platform[5] also do not prohibit the usage of publicly available data for research purposes. Extensive measures were taken to safeguard the privacy of all patients involved. No personal information outside of the CHQs was collected. In addition to automated measures, the dataset was manually inspected to ensure no personally identifiable information was present.

The individuals involved in annotating the dataset were provided monetary compensation for their work, which is above the minimum wage. The annotation process has also been anonymized to prevent any violations of the privacy of the annotators.

Acknowledgement

We are grateful to Islamic University of Technology (IUT) for providing the necessary funding for this research. We would also like to express our utmost gratitude to Dr. Rubaiya Bari for her professional input while preparing our dataset, as well as our team of annotators. This work would not have been possible without them.

[5] https://daktarbhai.com/

References

Asma Ben Abacha and Dina Demner-Fushman. 2019a. On the role of question summarization and information source restriction in consumer health question answering. *AMIA Summits on Translational Science Proceedings*, 2019:117.

Asma Ben Abacha and Dina Demner-Fushman. 2019b. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234.

Firoj Alam, Arid Hasan, Tanvirul Alam, Akib Khan, Janntatul Tajrin, Naira Khan, and Shammur Absar Chowdhury. 2021. A review of bangla natural language processing tasks and the utility of transformer models. *arXiv preprint arXiv:2107.03844*.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2022. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, and Rifat Shahriyar. 2023. 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*, pages 714–723.

Samit Bhattacharya, Monojit Choudhury, Sudeshna Sarkar, and Anupam Basu. 2005. Inflectional morphology synthesis for bengali noun, pronoun and verb systems. In *Proc. of the National Conference on Computer Processing of Bangla (NCCPB 05)*, pages 34–43.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *arXiv preprint arXiv:2310.04270*.

Md Tahmid Rahman Laskar, Enamul Hoque, and Xiangji Huang. 2022. Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, 48(2):279–320.

Md Tahmid Rahman Laskar, Jimmy Xiangji Huang, and Enamul Hoque. 2020. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5505–5514, Marseille, France. European Language Resources Association.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xiaojuan Ma, Xinning Gui, Jiayue Fan, Mingqian Zhao, Yunan Chen, and Kai Zheng. 2018. Professional medical advice at your fingertips: An empirical study of an online "ask the doctor" platform. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. *Mining text data*, pages 43–76.

Shaharior Rahman Razu, Tasnuva Yasmin, Taimia Binte Arif, Md Shahin Islam, Sheikh Mohammed Shariful Islam, Hailay Abrha Gesesew, and Paul Ward. 2021. Challenges faced by healthcare professionals during the covid-19 pandemic: a qualitative inquiry from bangladesh. *Frontiers in public health*, page 1024.

Kirk Roberts and Dina Demner-Fushman. 2016. Interactive use of online health resources: a comparison of consumer and professional questions. *Journal of the American Medical Informatics Association*, 23(4):802–811.

Syed Mohammad Shahed. 1993. Bengali folk rhymes: An introduction. *Asian folklore studies*, pages 143–160.

Manjira Sinha, Tirthankar Dasgupta, and Anupam Basu. 2016. Effect of syntactic features in bangla sentence comprehension. In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 275–284.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2022. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 3(4):100488.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2021. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 249–255.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2022a. Question-aware transformer models for consumer health question summarization. *Journal of Biomedical Informatics*, 128:104040.

Shweta Yadav, Deepak Gupta, and Dina Demner-Fushman. 2022b. Chq-summ: A dataset for consumer healthcare question summarization. *arXiv preprint arXiv:2206.06581*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Annotation Guidelines

The annotators were instructed to make the questions as short as possible while ensuring that no information required to answer the question was discarded. Aside from this, they were also provided with a list of examples to serve as a guideline in their work. The examples, provided in Table 5, cover perfect, passable, and poor summaries as approved by a practising physician.

A.2 Summary Annotation Samples

A few samples of the annotated summaries, along with their reference questions from the dataset, are provided in Table 6.

Sample 1	
Question	আমার বড় আপার সমস্যা। বয়স ৩২, ডায়াবেটিস রোগী। সর্বশেষ ডায়াবেটিস পরীক্ষা করিয়েছিলেন সপ্তাহখানেক আগে। সুগার লেভেল ছিল ১০। গতকিছুদিন আগে উনার দাঁতের গোড়া ফুলে উঠেছিল। এখন মনে হচ্ছে ফুলে উঠা জায়গাটা পেকে গিয়েছে, শুকাচ্ছে না। লকডাউন পরিস্থিতির কারণে ডাক্তার দেখানোও সম্ভব হচ্ছে না। উনার ছোট একটা বাচ্চা আছে। বর্ণিত অবস্থায় কি চিকি□সা নেওয়া প্রয়োজন জানালে খুব উপকৃত হবো।
Summary	বয়স ৩২। ডায়াবেটিস আছে। সপ্তাহখানেক আগে ডায়াবেটিস পরীক্ষা করলে সুগার লেভেল ১০ হয়। দাঁতের গোড়া ফুলে পেকে গেছে, শুকাচ্ছে না। পরামর্শ চাই।
Analysis	**Perfect Summary**. The summary specifies the age of the patient (relevant to diabetes), the fact that they have diabetes, the sugar level during the last test as well as the issue the patient is currently facing. All the unnecessary information has been successfully removed, such as the relationship with the patient, the fact that they cannot visit a doctor due to the lockdown and that the patient has a child.
Sample 2	
Question	ডাক্টার বলছে রুট ক্যানেল করতে। রুট ক্যানেল করলে ভালো নাকি দাত ফেলে দিয়ে দাত লাগালে ভালো? ফিউচার এর জন্য কোনটা বেটার হবে। দাত ফেলতে ভয় পাচ্ছে চোখ বা মাথা ব্যাথার জন্য। আর রুট ক্যানেল করলেও নাকি কয়েকমাস পর ব্যাথা হয় দাতে। সাজেশন চাচ্ছি একটু কি করলে ভালো হয়। দাত ব্যাথায় টিকতে পারছে না।
Summary	প্রচুর দাঁত ব্যথা। ডাক্তার রুট ক্যানাল করতে বলেছে।
Analysis	**Poor Summary**. The summary does a poor job of retaining the actual questions the patient had. The patient wanted the doctor's opinions on various things such as whether to do a root canal or remove the tooth entirely and whether doing a root canal will cause pain after a few months.
Sample 3	
Question	আসসালামুয়ালাইকুম, সপ্তাহ খানেক আগে শীলা বৃষ্টিতে ভিজেছি। এরপর ৪ - ৫ চামচ আইস ক্রিম খেয়েছিলাম। ২৫ তারিখ থেকেই শরীরের অবস্থা ভালো মনে হচ্ছিল না। ২৬ তারিখ মাগরিবের পর জ্বর আসে। পরিমাপ করে দেখা যায় ১০২ ডিগ্রি। পরদিন জ্বর কমে যায়, কিন্তু অসহ্য রকম গলা ব্যাথা শুরু হয়, যা এখন পর্যন্ত আছে। ঢোক গেলা যাচ্ছে না। যেসকল ঔষধ খেয়েছি: 1, নাপা এক্সটেন্ড ট্যাব, খাওয়া শেষ 2, ফেকযো ট্যাব, খাওয়া শেষ 3, রেলিজিন ট্যাব, খাওয়া শেষ 4, মিউকলিট সিরাপ, অল্প বাকি 5, জি ম্যাক্স ট্যাব, খাওয়া শেষ 6, ভিকজিন, চলে
Summary	বৃষ্টিতে ভিজে জোর ১০২। ঔষধ খেয়েছি, নাপা এক্সটেন্ড ট্যাব,, ফেকযো ট্যাব, রেলিজিন ট্যাব, জি ম্যাক্স ট্যাব, খাওয়া শেষ। মিউকলিট সিরাপ, অল্প বাকি, লিভোকজিন, চলে।?
Analysis	**Passable Summary**. The summary accurately captures a large amount of information, but makes a critical mistake. The patient mentioned that their fever has decreased and that they are suffering from a throat ache now. The summary does not mention this.

Table 5: Examples used as annotation guidelines.

A.3 Experimental Setup

The experimental setup consisted of an Nvidia 3090 GPU with 24 GB of VRAM. The Trainer library, available through Hugging Face was utilized, along with CUDA Version 11.6. The dataset was divided into training, validation, and test sets using the split ratio $80 : 10 : 10$. The models were trained for 50 epochs using a cross-entropy loss function along with the AdamW optimizer. Input sequences were truncated to a maximum length of 512 tokens, and the output sequences were limited to 128 tokens. Other hyperparameters include a batch size of 16, a weight decay of 0.03, and a learning rate of 1e-4 used with a linear learning rate scheduler.

Original Question	Annotated Summary
আমাকে প্রশ্নের উত্তর দেয়া হয়েছে, এ জন্য আপনাদের অসংখ্য ধন্যবাদ জানাচ্ছি, বিষয়টি হল আমি একজন ডাক্তারের দেয়া ব্যবস্থাপত্র এই এ্যাপে আপলোড করেছি, এবং এই ব্যবস্থা পত্র অনুযায়ী এখন ও ঔষধ গ্রহন করছি, সে ক্ষেত্রে ১। নিউরো বি খাওয়া হলে কোন অসুবিধা হতে পারে কিনা। ২। ঘুমের সমস্যার জন্য রি লাইফ ট্যাবলেট খাওয়া যায় কিনা , কারন ঔষধটি বাসায় সংরক্ষিত আছে। দয়া করে ব্যবস্থা পত্র দিবেন।	নিউরো বি খেলে অসুবিধা হবে কি না এবং ঘুমের জন্যে রি লাইফ ট্যাবলেট খাওয়া যায় কি না?
আসসালামু আলাইকুম। স্যার অনেক দিন ধরে আমার মাথায় ও সরিরে চুলকানি। মাথা ও দারির ভেতরে ঘাও দিয়ে ভরে গেছে। ঘাও এর কারণ এ মাথার চুল ও পরে যাচ্ছে। আমি অনেক ওষুধ খেয়েছি কোন কাজ হয়নি। দয়া করে বলবেন কি ওষুধ খেলে ভালো হবে।????	অনেক দিন ধরে মাথায় ও শরীরে চুলকানি। মাথা ও দাড়ির ভেতরে ঘা, মাথার চুল পরে যাচ্ছে। কি করবো?
হ্যালো আসসালামুআলাইকুম, হার্টে কোলেস্টরল এর মাত্রা কিভাবে নিয়ন্ত্রণে রাখতে পারবো, কোন মেডিসিন গ্রহন করলে উপকার পাবো, দয়া করে একটু জানাবেন? এবং উচ্চ রক্তচাপ নিয়ন্ত্রণে রাখতে কোন ধরনের খাবার খাবো, এবং কোন ধরনের খাবার বর্জন করবো, সে ব্যাপারে একটু জানাবেন!!	কোলেস্টরল এবং উচ্চ রক্তচাপ নিয়ন্ত্রণ করতে কি ঔষুধ খাব এবং কি খাবার বর্জন করব?

Table 6: Samples of annotated summaries from the BanglaCHQ-Summ dataset