# Answer-agnostic Bangla Question-answer Pair Generation Using Transformer-based Approaches

**Authors**

Md Sajid Altaf (180041203)

Syed Mohammed Sartaj Ekram (180041204)

Adham Arik Rahman (180041219)

**Supervisor**

Dr. Md. Azam Hossain

Assistant Professor

Department of Computer Science and Engineering (CSE)

Islamic University of Technology (IUT)

**A thesis submitted to the Department of CSE**

**in partial fulfillment of the requirements for the degree of**

**Bachelor of Science in CSE**


**Department of Computer Science and Engineering (CSE)**

**Islamic University of Technology (IUT)**

**Organization of the Islamic Cooperation (OIC)**

**Gazipur, Bangladesh**

**May 2023**

## Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by Md Sajid Altaf, Syed Mohammed Sartaj Ekram and Adham Arik Rahman under the supervision of Dr. Md. Azam Hossain, Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

***Authors:***

27/05/23

Md Sajid Altaf

Student ID: 180041203

27/05/23

Syed Mohammed Sartaj Ekram

Student ID: 180041204

27/05/23

Adham Arik Rahman

Student ID: 180041219

***Supervisor:***

২৭/০৫/২০২৩

Dr. Md. Azam Hossain

Assistant Professor,

Department of Computer Science and Engineering

Islamic University of Technology (IUT)

# Abstract

High-resource languages, such as English, have access to a plethora of datasets with various question-answer types resembling real-world reading comprehension. However, there is a severe lack of diverse and comprehensive question-answering datasets in under-resourced languages like Bangla. The ones available are either translated versions of English datasets with a niche answer format or created by human annotations focusing on a specific domain, question type, or answer type. To address these limitations, we introduce BanglaRQA, a reading comprehension-based Bangla question-answering dataset with various question-answer types. BanglaRQA consists of 3,000 context passages and 14,889 question-answer pairs created from those passages. The dataset comprises answerable and unanswerable questions covering four unique categories of questions and three types of answers. In addition, we also implemented four different Transformer models for question-answering on the proposed dataset. The best-performing model achieved an overall 62.42% EM and 78.11% F1 score. However, detailed analyses showed that the performance varies across question-answer types, leaving room for substantial improvement of the model performance. Furthermore, we demonstrated the effectiveness of BanglaRQA as a training resource by showing strong results on the bn_squad dataset.

We focus on Bangla Question-answer pair generation for the next part of our work. Bangla, being a less explored language in NLP, lacks comprehensive research in the domain of question-answer pair generation. We focus on developing this untapped sector by fine-tuning BanglaT5, a generative model on the BanglaRQA dataset. The quality of the generated questions is first evaluated using various metrics. The best-performing model, BanglaT5, achieved a BLEU score of 21.56 and a BERT score of 85.04, indicating that the generated questions exhibit decent quality. Subsequently, the research progresses toward the main task of generating question-answer pairs. The quality of the generated pairs is evaluated through human assessment and baseline comparison, demonstrating that the generated QA pairs possess comparable quality to human-annotated QA pairs. Therefore, this work proposes an end-to-end Question-Answer-Generation (QAG) pipeline and presents a reading-comprehension-based dataset, that has the potential to contribute to future research.

# Acknowledgement

We are delighted to present our thesis work, marking a significant milestone in our Bachelor of Science journey. As we embark on this auspicious moment, we would like to express our profound gratitude to Almighty Allah for the countless blessings that have guided us throughout this research endeavor, enabling us to achieve success and completion.

We extend our deepest appreciation to Dr. Md. Azam Hossain, Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology (IUT), for his exceptional role as our adviser and mentor. His unwavering support, motivation, and insightful suggestions have been truly invaluable throughout the journey of this thesis. We acknowledge that without his guidance, our research would not have followed a well-defined path in the realm of academia. His invaluable expertise, dedicated time, and meaningful contributions have been instrumental from the very inception of our thesis. We are sincerely grateful for his consistent and enthusiastic guidance, as well as his precious advice that has greatly shaped and enriched our thesis work.

We would also like to thank Mr. Mohammed Saidul Islam and Md. Mezbaur Rahman, Lecturer, Department of Computer Science and Engineering, Islamic University of Technology for their valuable advice and contributions.

We would like to extend our vote of thanks to all the respected jury members of our thesis committee for their insightful comments and constructive criticism of our research work. Surely they have helped us to improve this research work. We are truly grateful for their time, commitment, and valuable contributions, which have undoubtedly enhanced the quality and depth of our thesis.

Last but certainly not least, we would like to extend our heartfelt appreciation to all the esteemed faculty members of the Computer Science and Engineering department at the Islamic University of Technology. Their unwavering support and guidance have played a crucial role in creating a conducive and pleasant working environment for us. Their expertise and willingness to support have been invaluable resources that we have relied upon. We sincerely thank them for their continuous encouragement and for fostering an atmosphere of learning and growth within the department.

# Contents

# Chapter 1

# Introduction

## 1.1 Background Study

Reading comprehension is the ability to process the information gained from reading a context passage, comprehend the meaning of both the context passage and the question asked, and then respond based on the reader's understanding and knowledge of the topic [1]. This process also involves determining which questions cannot be answered based on the context. Therefore, reading comprehension is widely recognized as a crucial test for evaluating humans' and machines' natural language comprehension.

Reading comprehension-based question-answering generally comprises three parts: context, question, and answer. Each of them has a wide range of types and formats. In the case of the English language, a large number of distinct datasets based on reading comprehension have been developed in order to capture the diversity of these components. For example, datasets in English are accessible with a single document as context [2], multiple documents as context [3], fill in the blank type questions [4], questions in natural language [5], single span-based replies [2], answers in natural language [6], and so forth.

Question-answer pair generation is a widely researched topic in natural language processing. It involves generating question-and-answer pairs from a given context without human involvement. It has become increasingly important in recent years as it can alleviate the laborious task of manually forming question-answers. By learning to generate question-answer

pairs on a larger scale in a shorter time, these systems can help assess a user's comprehension of a piece of text. The task has a wide range of practical applications in various sectors, such as education [7], medical science [8], language learning [9], etc. Chatbots, search engines, language learning tools [10], FAQ generation, intelligent tutoring systems, and virtual assistants [9] are just a few examples of how this technology can be applied to improve user experiences and provide valuable insights.

Question Answering Systems has been around since the 1960s and the first system was built using a set of manual rules. Since then, there have been significant advancements in the architecture of these systems, with changes ranging from Recurrent Neural Networks (RNNs) [11] to transformer-based models [12]. Similarly to QAS, the early research for automated Question Generation System relied heavily on manually crafted set of rules focused predominantly on the syntactic structure of the text [13, 14]. Down the line, with the increasing success of seq-to-seq learning models, researchers have been able to move beyond relying on a fixed set of rules for generating questions [15, 16]. In recent years, transformer models such as BERT [17], GTP [18], XLNet [10], etc. have been used to pre-train on large corpora of data, leading to significant improvements in natural language generation tasks [12]. These improvements in Question Answering Systems and Question Generation Systems further helped to pave the way for the development of Question-answer pair generation systems.

Generating high-quality question-answer pairs from unstructured text is still quite challenging. The majority of existing approaches deal with answer-aware question generation, where the model is fed with an answer chunk and its context to produce the question [10]. These often require the explicit selection of an answer span in the input context, which is usually done by highlighting specific tokens. However, this approach can add significant overhead and is not ideal in situations where a clear list of key terms is not available [19]. In real-world applications, answers aren't always known beforehand. Answer-agnostic question generation takes this into consideration and enables the model to generate questions without any prior knowledge of the corresponding answers. This increases the model's degree of freedom resulting in a more diverse set of questions including unanswerable questions [20]. Answer-agnostic question generation systems accelerated the development of auto-

matic question-answer pair generation systems as it only makes sense to find corresponding answers once we have already generated questions.

## 1.2 Motivation

Bangla is the world's seventh most frequently spoken language, as over 230 million people speak it in Bangladesh and India [21]. Although Bangla is a rich and diverse language, it is severely under-resourced for natural language processing. This is mostly attributed to the scarcity of necessary resources, such as labeled datasets, language models, and effective machine learning (ML) techniques for a variety of different NLP applications. A few datasets are available for reading comprehension-based question-answering in Bangla. These datasets, however, are either translated versions of the English datasets [22, 23], or very small datasets that focus on a specific topic area, such as general knowledge [24] or only for answer type [25]. This illustrates the need for a diverse and high-quality dataset for NLP research in Bangla for question-answer based on reading comprehension.

While there have been at least some works focused on reading comprehension-based question-answering introducing different datasets and language models, little to no attention has been given to question-answer pair generation. With the rapid advancements in natural language processing tasks, English language has seen tremendous progress with the use of different generative models in this arena [26]. However, it is imperative that under-resourced languages also make progress in this field, to ensure equitable access to the benefits of such advancements.

## 1.3 Problem Statement

We worked on BanglaRQA, a benchmark dataset for Bangla question-answering based on reading comprehension that contains a wide variety of question-answer types. This dataset comprises 3000 context passages covering a wide range of domains with 14,889 question-answer pairs. Out of the 14,889 questions, 3,631 questions were unanswerable from their

respective passages. The unanswerable questions are constructed in such a way that they seem pertinent to the passages to which they belong. This mix of answerable and unanswerable questions trains language models when to answer and when not to respond, resulting in improved linguistic ability. Furthermore, the dataset includes a wide variety of question types, and these question types are separated into four categories, as illustrated in Figure 1.1: factoid, causal, confirmation, and list. Consequently, it guarantees that the dataset contains different challenges to answering different types of questions.

For the answerable questions, the answers can be classified into one of three groups: single span, multiple spans, or yes/no, covering the extractive question-answering domain. Multiple-span answers enable information to be accumulated from different parts of the context passage as the answer. Yes/No answers require inference skills based on the passage's context making the dataset more robust.

To estimate the difficulty of BanglaRQA, this study also fine-tuned four different pre-trained Transformers models, namely, BanglaT5 [27], mT5 [28], BanglaBERT [23], mBERT [29]. BanglaT5, the best-performing model, achieved an average of 62.42% EM and 78.11% F1 score on the test set. However, the EM and F1 scores were lower for some specific question-answer types, indicating some of our dataset's challenges.

After training BanglaT5 on our dataset, we tested it on the previously available bn_squad dataset (a translated version of SQuAD 2.0) [23], yielding 70.20% EM and 75.79% F1 score. Previously [27], when BanglaT5 was trained and tested on the bn_squad dataset, it yielded 68.49% EM and 74.77% F1 score. This proves that BanglaRQA is a valuable resource for training language models by demonstrating the model's capacity to generalize to bn_squad.

For the question-answer pair generation part, we fine-tuned various generative models e.g. mT5 and BanglaT5 on BanglaRQA dataset to produce Bangla question-answer pairs. We employed an answer-agnostic approach for the question generation system, and subsequently, we utilized the question-answering system to find the answers from the given context and generated questions. In the evaluation phase, we utilized various metrics such as EM, F1, BLEU, and BERT to estimate the quality of the generated questions. Finally, to assess the overall quality of the generated question-answer pairs we do a human assessment and base-

line comparison. While doing the human assessment, the annotators took a subset of the generated question-answer pairs and identified different error types. As human assessment is not feasible for all synthetic pairs, we do a baseline comparison where we combined the generated QA pairs with the original BanglaRQA dataset and trained our models on the augmented dataset. We then measured the performance of our models for Question-Answering on this augmented dataset and found out that we achieve comparable performance to the baseline model. These results suggest that the quality of the generated QA pairs is similar to that of human-annotated QA pairs.

**Contributions**. Our contributions are the followings:

- We present BanglaRQA, a human-annotated dataset for Bangla reading comprehension containing 14,889 question-answer pairs curated from 3000 passages.
- The proposed dataset contains a variety of question types, including factoids, causal, confirmation, and list questions. In addition, both answerable and unanswerable question-answer pairs are included.
- Proposed BanglaRQA additionally includes answers that are divided into three categories: single span, multiple spans, and yes/no, encompassing the domain of extractive question-answering.
- We fine-tuned four different Transformer models: BanglaT5, mT5, BanglaBERT, and mBERT to establish baseline performances on the proposed dataset. Furthermore, we analyzed the performance of BanglaT5, the best performer in our dataset, on various question-answer types.
- We demonstrate that BanglaRQA can be a resourceful dataset to train language models by showing its generalization capability to bn_squad.
- We fine-tuned BanglaT5 on BanglaRQA dataset to produce Bangla question-answer pairs.
- We assessed the quality of generated questions based on different metrics.

- We proved that generated question-answer pairs are similar to human-annotated QA pairs based on human assessment and baseline comparison.

**a) Factoid question with Single Span answer**

**Context:** অপবিজ্ঞান বা সিউডোসায়েন্স একটি দাবি, বিশ্বাস বা অনুশীলন যা বিজ্ঞান হিসাবে উপস্থাপিত হয়, তবে যা বৈজ্ঞানিক পদ্ধতি অনুসরণ করে না। যদি গবেষণার কোনও বিষয়কে বৈজ্ঞানিক পদ্ধতির মানদণ্ড অনুসারে উপস্থাপন করা হয় তবে এটি এই মানদণ্ডগুলি অনুসরণ করে না। অপবিজ্ঞান ক্ষতিকারক হতে পারে যেমনঃ অ্যান্টি-ভ্যাকসিনকর্মীরা অপবৈজ্ঞানীক গবেষণা উপস্থাপন করে, যা ভ্যাকসিনগুলির সুরক্ষাকে অন্যায়ভাবে প্রশ্নবিদ্ধ করে। কোনও প্রামান ছাড়াই হোমি ...

**Question:** অপবিজ্ঞানের অপর নাম কী?
**Question Type:** factoid
**Is Answerable:** yes
**Answer:** সিউডোসায়েন্স
**Answer Type:** single span

**b) Confirmation question with Yes/No answer**

**Context:** তথ্য গোপণীয়তা বা ডাটা গোপণীয়তা বা ডাটা সুরক্ষা হল ডাটা, প্রযুক্তি, জনগণের গোপনীয়তার প্রত্যাশা এবং আইন সংক্রান্ত ও রাজনৈতিক বিষয়াদির সংগ্রহ এবং বিতরণের মধ্যকার সম্পর্ক।
গোপনীয়তা যেখানে ব্যক্তিগত চিহ্নিতকরন তথ্য বা অন্যান্য স্পর্ষকাতর তথ্য সংগ্রহ এবং জমা হয় (ডিজিটালভাবে বা অন্যকোন ভাবে) সেখানেই সম্পর্কযুক্ত। অনুপযুক্ত, অকার্যকর ...

**Question:** তথ্য গোপনীয়তা কি ডিজিটালভাবে জমা হওয়া স্পর্শকাতর তথ্যের সাথে সম্পর্কিত?
**Question Type:** confirmation
**Is Answerable:** yes
**Answer:** হ্যাঁ
**Answer Type:** yes/no

**c) List question with Multiple Spans answer**

**Context:** দ্য ডার্ক নাইট ২০০৮ সালে মুক্তি পাওয়া ক্রিস্টোফার নোলানের পরিচালিত একটি মার্কিন সুপারহিরো চলচ্চিত্র। ডিসি কমিকস এর সুপারহিরো ব্যাটম্যানকে নিয়ে নির্মিত এই চলচ্চিত্র ২০০৫ সালের ব্যাটম্যান বিগিনস চলচ্চিত্রের সিক্যুয়েল। এতে ব্যাটম্যান চরিত্রে অভিনয় করেন ব্রিটিশ অভিনেতা ক্রিশ্চিয়ান বেল। অন্যান্য অভিনয় শিল্পীদের মধ্যে ছিলেন মাইকেল কেইন, হিথ লেজার, গ্যারি ওল্ডম্যান, অ্যারন একহার্ট, ম্যাগি জিলেনহল ও মরগান ফ্রিম্যান।

চলচ্চিত্রটি নির্মিত হয় যুক্তরাষ্ট্র ও যুক্তরাজ্যের যৌথ প্রযোজনায়। ...

**Question:** দ্য ডার্ক নাইটে কে কে অভিনয় করেন?
**Question Type:** list
**Is Answerable:** yes
**Answer:** ক্রিশ্চিয়ান বেল; মাইকেল কেইন; হিথ লেজার; গ্যারি ওল্ডম্যান; অ্যারন একহার্ট; ম্যাগি জিলেনহল; মরগান ফ্রিম্যান
**Answer Type:** multiple spans

**d) Causal question with Single Span answer**

**Context:** নিউটন ট্রিনিটি কলেজ থেকে ১৬৬১ সনে মেট্রিকুলেশন পাশ করেন। কলেজে অধ্যয়নকালে তিনি তার পড়াশোনার খরচ চালানোর জন্য কলেজের বিভিন্ন স্থানে ভৃত্যের কাজ করতেন। ছাত্র হিসেবে বড় কোন কিছু তিনি করেছেন বলে ট্রিনিটি কলেজের কোন দলিলপত্র লেখা নেই। তবে জানা যায় তিনি মূলত গণিত ও বলবিজ্ঞান বিষয়ে অধিক পড়াশোনা করেছিলেন। ট্রিনিটি কলেজে প্রথমে তিনি কেপলারের আলোকবিজ্ঞান বিষয়ক সূত্রের উপর অধ্যয়ন করেন। এরপর অবশ্য তিনি ইউক্লিডের জ্যামিতির প্রতি মনোনিবেশ করেন। কারণ মেলা থেকে কেনা জ্যোতিষ শাস্ত্রের একটি বইয়ে উল্লেখিত বেশ কিছু রেখাচিত্র তিনি বুঝতে পারছিলেন না। এগুলো বোঝার জন্য ইউক্লিডের জ্যামিতি জানা থাকাটা আবশ্যিক ছিল। তা সত্ত্বেও নিউটন বইটির ...

**Question:** কেন নিউটন কলেজের বিভিন্ন স্থানে ভৃত্যের কাজ করতেন?
**Question Type:** causal
**Is Answerable:** yes
**Answer:** পড়াশোনার খরচ চালানোর জন্য
**Answer Type:** single span

**e) Unanswerable Factoid question**

**Context:** আইনস্টাইন পদার্থবিজ্ঞানের বিভিন্ন ক্ষেত্রে প্রচুর গবেষণা করেছেন এবং নতুন উদ্ভাবন ও আবিষ্কারে তার অবদান অনেক। সবচেয়ে বিখ্যাত অবদান আপেক্ষিকতার বিশেষ তত্ত্ব (যা বলবিজ্ঞান ও তড়িচ্চৌম্বকত্বকে একীভূত করেছিল) এবং আপেক্ষিকতার সাধারণ তত্ত্ব (যা অসম গতির ক্ষেত্রে আপেক্ষিকতার তত্ত্ব প্রয়োগের মাধ্যমে একটি নতুন মহাকর্ষ তত্ত্ব প্রতিষ্ঠিত করেছিল)। তার অন্যান্য অবদানের মধ্যে রয়েছে আপেক্ষিকতাভিত্তিক বিশ্বতত্ত্ব, কৈশিক ক্রিয়া, ক্রান্তিক উপলবৎ বর্ণময়তা, পরিসংখ্যানিক বলবিজ্ঞান ও কোয়ান্টাম তত্ত্বের বিভিন্ন সমস্যার সমাধান যা তাকে অণুর ব্রাউনীয় গতি ব্যাখ্যা করার দিকে পরিচালিত করেছিল, আণবিক ক্রান্তিকের সম্ভাব্যতা, এক-আণবিক গ্যাসের কোয়ান্টাম তত্ত্ব, নিম্ন বিকরণ ঘনত্বে আলোর তাপীয় ধর্ম (বিকিরণের একটি তত্ত্ব যা ফোটন তত্ত্বের ভিত্তি রচনা করেছিল), একীভূত ক্ষেত্র তত্ত্বের প্রথম ধারণা দিয়েছিলেন এবং পদার্থবিজ্ঞানের জ্যামিতিকীকরণ করেছিলেন।

**Question**: আপেক্ষিকতাভিত্তিক বিশ্বতত্ত্ব কী বোঝায়?
**Question Type:** factoid
**Is Answerable:** no
**Answer:**
**Answer Type:**

**Figure 1.1:** Samples of different question-answer types of BanglaRQA with truncated passages where: a) Factoid question with Single Span answer. b) Confirmation question with Yes/No answer. c) List question Multiple Spans answer. d) Causal question with Single Span answer. e) Unanswerable Factoid question.

# Chapter 2

# Related Work

This section presents existing works for reading comprehension-based question-answering datasets in English and Bangla and subsequently on question-answer pair generation.

## 2.1 English Reading Comprehension Datasets

SQuAD 2.0 [2] and NewsQA [5] are large-scale human-annotated datasets with single document as context passage. Their questions are in natural language, including unanswerable ones, and the answers are single-span based. On the other hand, QAngaroo [3], and HotpotQA [30] are datasets where the questions require finding and reasoning over multiple supporting documents to answer.

ReClor dataset [31] is collected from exams like GMAT and LSAT, making it very challenging. Another of this kind is RACE dataset [32], collected from middle school and high school English examinations in China.

MS MARCO [6] is a large-scale dataset where the passages are collected from web documents, the questions are collected from Bing search queries, and the answers are human-generated in natural language. A question in the MS MARCO dataset may have multiple or no answers.

Natural Questions [33] dataset comprises both answerable and unanswerable questions searched by real users in the Google search engine. The context of each question here is an entire Wikipedia article. For the answerable questions, the answers can be either a long extracted paragraph from the context or a short answer containing one or two entities.

MultiRC [34] is a dataset where questions contain multiple sentences and can be answered from their corresponding passages. The answers need not be only span based. CoQA [35] is a large-scale dataset where each sample is a question-answer conversation between two crowd-workers about a context passage. Another conversational dataset is QuAC [36], where in a sample, a student asks a question about a hidden Wikipedia passage, and a teacher answers with spans from that passage.

## 2.2 Bangla Reading Comprehension Datasets

Some of the recent works [22, 23] include creating translated Bangla datasets from SQuAD 2.0 [2]. Their passages cover a wide range of topics. They include answerable and unanswerable questions. Answerable questions are answered by only a single span from their respective passage.

A very small dataset focusing on the specific domain of general knowledge [24] was also developed by using sources like Facebook and Google. Another dataset on factoid question answering [37] was developed. The whole dataset consisted of 1,676 paragraphs, from which 8,027 question-answer pairs were generated. On average, the length of each question consisted of 10-11 words, and the answer consisted of 3-4 words. BQuAD [25] dataset consists of a collection of question-answer pairings and contexts from a variety of fields. They used Bangla Wikipedia articles as their source. The dataset comprises 5000 context paragraphs, each with 2-5 questions. There were 13 thousand question-answer pairings in total. Even though this dataset had no restrictions on the type of questions, the question-answer format was exactly like SQuAD 1.1 [38] with no unanswerable questions and answers of only a single span from the context passage.

Figure 2.1 illustrates a comparison among BanglaRQA and previously available Bangla question-answering datasets.

| **Dataset** | **Curation Process** | **Unanswerable Questions?** | **List Questions?** | **Confirmation (Yes/No) Question-Answers?** | **Multiple Span Answers?** |
|---|---|---|---|---|---|
| Bengali-SQuAD | T | ✔ | X | X | X |
| bn_squad | T | ✔ | X | X | X |
| General Knowledge | **HA** | X | X | X | X |
| Factoid QA | **HA** | X | X | X | X |
| BQuAD | **HA** | X | X | X | X |
| **BanglaRQA** | **HA** | ✔ | ✔ | ✔ | ✔ |

**Figure 2.1:** Comparison among BanglaRQA and previously available Bangla reading comprehension datasets. Here, 'T' = Translation, 'HA' = Human Annotation.

## 2.3 Question-answer Pair Generation

The task of producing question-answer pairs given a context has been the subject of numerous works. While there are several publications available on English question-answer pair generation, the same cannot be said for the Bangla language. The following pieces discuss some of the key studies on question-answer pair formation.

### 2.3.1 Self-Attention Architectures for Answer-Agnostic Neural Question Generation

The task of producing questions given a context or passage is the main topic of the work titled "Self-Attention Architectures for Answer-Agnostic Neural Question Generation" [39]. The research suggests a unique method for question generation that is answer-neutral and based on self-attention architectures. The authors stress the value of answer-agnostic question generation, in which the model generates questions without considering particular answers. Because of this method, the model can handle a wide range of situations and answer types. The suggested paradigm makes use of self-attention mechanisms built into a transformer-based architecture. The model may generate questions while attending to various sections of the input text thanks to self-attention. The model can produce more cogent and pertinent queries by accurately collecting contextual data.On benchmark datasets, the authors test various iterations of the suggested self-attention architecture and evaluate how well they perform in comparison to currently used question creation techniques. They assess the questions produced using accepted measures like BLEU and ROUGE.

The outcomes show that the suggested self-attention structures perform better than earlier approaches, reaching cutting-edge results in answer-agnostic question generating tasks. The trials also show how well self-attention works for gleaning pertinent information from the environment and formulating insightful queries. Overall, the research proposes a novel method for creating open-ended questions utilizing self-attention architectures. The findings show that self-attention processes have the ability to raise the caliber and relevance of inquiries that are created.

Limitations of this study exist. It might not be compared to other approaches for creating questions, which would make it harder to evaluate its effectiveness and benefits. The evaluation could be restricted to a small number of datasets, which would reduce the generalizability of the suggested structures. The quality and diversity of the questions that were generated may not have been sufficiently analyzed, and the paper may have overlooked the difficulty of coming up with questions that depended on the answer. Additionally, the investigation of hyperparameters, model architectures, and alterations in the self-attention mechanism may be limited, which limits knowledge of how they affect the effectiveness of question creation.

### 2.3.2 Automatic Question Generation from Sentences

The paper [40] discusses the challenges posed by Question Generation (QG) and Question Answering (QA) in natural language understanding and interfaces. It suggests that automated QG systems have the potential to assist humans in asking good questions and fulfilling their inquiry needs effectively. The focus of the paper is on the task of Sentence-to-Question generation, where a given sentence is used as input to generate a set of questions that the sentence either contains, implies, or requires answers to. The authors propose an approach that involves breaking down complex sentences into elementary sentences using a syntactic parser. Additionally, a named entity recognizer and a part of speech tagger are applied to encode essential information for each of these sentences. To determine the possible types of questions to be generated, the sentences are classified based on their subject, verb, object, and preposition. This classification helps guide the question generation process. For the experiments and evaluation, the authors utilize the TREC-2007 (Question Answering Track) dataset. They likely use this dataset to assess the performance and effectiveness of their proposed approach.

In summary, the paper focuses on automating the task of generating questions from sentences. The authors present an approach that involves breaking down complex sentences, encoding necessary information, classifying sentences based on linguistic components, and utilizing a specific dataset for evaluation purposes.

The paper has several limitations. It may have a limited evaluation on diverse datasets, hin-

dering the generalizability of the proposed approach. There might be a lack of comparison with alternative methods, making it difficult to assess the advantages of the proposed method. The paper may not thoroughly analyze the quality of generated questions, neglect the importance of context in question generation, and insufficiently explore linguistic variations. Additionally, the paper may not address scalability and efficiency concerns when dealing with large amounts of data.

### 2.3.3 On the Generation of Medical Question-Answer Pairs

To evaluate the generated questions, the authors employ human judges to assess the quality of the generated question-answer pairs. They also compare the generated pairs with human-created pairs to measure the performance of their approach. The experimental results show that the proposed method achieves promising results in generating medical question-answer pairs. The generated questions are deemed of high quality and are comparable to human-created pairs. The paper[8] highlights the importance of generating medical question-answer pairs and introduces a method that utilizes large-scale data from online health forums. The approach demonstrates the effectiveness of using supervised learning with a sequence-to-sequence model to generate relevant and coherent questions in the medical domain.
The authors use human judges to analyze the generated questions and evaluate the resulting question-answer combinations. In order to gauge the effectiveness of their strategy, they also contrast the generated pairs with pairs made by humans. The experimental findings demonstrate that the suggested approach generates medical question-answer pairs with promising results. The generated questions are thought to be of a high caliber and are comparable to pairs made by humans. In addition to highlighting the significance of creating medical question-answer pairings, the research also proposes a technique that makes use of extensive data from online health forums. The method shows how well supervised learning combined with a sequence-to-sequence model may produce pertinent and cogent queries in the field of medicine.
The sole goal of the research is to develop new question-answer pairs that will improve a QA model's performance on challenge sets. Although creating new pairs increases the diversity

of the training data and allows the model to handle a wider range of queries, it ignores other potential sources of error, such as insufficient or inaccurate context information or weaknesses in the model's reasoning skills. Furthermore, the work surpasses several baselines by using merely a VAE as the generative model. Investigating different generative models or data augmentation methods could make the model more resilient to challenging sets. SQuAD and Natural Questions, two commonly used QA datasets that reflect real-world problems, are utilized to evaluate the suggested approach. These datasets, however, have limits and might not fully account for all of the difficulties a QA model might run into. A thorough knowledge of the approach's effectiveness and generalizability would result from testing it on other datasets and problems.

There are a few issues with the study that should be mentioned. The performance of the approach on larger and more varied datasets cannot be fully understood because the evaluation was only done on a limited dataset of medical QAPs. Since the approach primarily depends on domain-specific knowledge, it is unclear whether it can be applied to areas other than medicine. It is difficult to duplicate or alter the strategy due to the rule-based component's lack of a comprehensive explanation. The coverage of prospective user enquiries is constrained by the restricted consideration of various query categories. Additionally, the article only evaluates the quality and accuracy of the generated answers, particularly focusing on question development.

### 2.3.4 Automatic Question-answer Pairs Generation and Question Similarity Mechanism in Question Answering System

The paper[12] presents a method that automatically generates question-answer pairs (QAPs) and incorporates a question similarity mechanism to enhance the effectiveness of a question answering (QA) system. The approach utilizes a deep learning model to generate QAPs based on a text corpus. Additionally, a similarity measure is employed to identify questions in the corpus that are similar to the user's query, enabling the retrieval of relevant answers. In summary, the paper introduces a technique for automated QAP generation and leverages question similarity to improve the performance of a QA system.

The paper presents a deep learning-based method for producing question-answer pairs (QAPs) from a corpus of text data, specifically using an LSTM network. In order to capture semantic similarity and guarantee high-quality results, the model is trained on a huge dataset of autonomously created QAPs.

The authors suggest a question similarity mechanism based on word embeddings and cosine similarity in order to improve the performance of the question answering (QA) system. This approach finds the question in the corpus that most closely resembles the supplied question by comparing it to previously created question embeddings. The system responds by retrieving the corresponding response and displaying it.

On two datasets, Yahoo! Answers and WikiQA, the effectiveness of the technique is assessed and contrasted with several industry standards. The outcomes demonstrate that the proposed technique outperforms the baselines on both datasets in terms of precision, recall, and F1 score. A user survey is also carried out, and users find the system's responses to be very pertinent and beneficial.

In conclusion, the research offers a fascinating method for automatically producing high-quality QAPs and utilizing a question similarity mechanism to improve QA system performance. On numerous datasets, the method is rigorously assessed and shows gains in system performance and user satisfaction.

There are several restrictions on the document that need to be acknowledged. Bias in the randomly produced question-answer pairs (QAPs) is not addressed. The research does not provide a full analysis of the proposed question similarity mechanism's effectiveness. In addition, the article focuses exclusively on technological issues without considering any ethical ramifications of employing a QA system to automatically generate and display answers to consumers.

### 2.3.5 Asking Questions the Human Way: Scalable Question-Answer Generation from Text Corpus

In this research [10], a unique method for producing high-quality question-answer pairs (QAPs) from a text corpus is presented. Using this method, questions are created based on

important phrases in the text, and then they are ranked using a scoring model that is based on deep learning.

The authors use part-of-speech tagging, named entity identification, dependency parsing, and other statistical and rule-based techniques to find crucial phrases. Then, while taking into account syntactic and semantic restrictions, these important phrases are employed to construct queries. The authors suggest using a convolutional neural network (CNN) scoring model to score the generated queries according to their quality.

The method is tested against numerous benchmarks and evaluated on three datasets: SQuAD, WikiQA, and Yahoo! Answers. The outcomes show that the suggested method outperforms the baselines in terms of the caliber and variety of the questions. According to user evaluations from user research, the system's queries are extremely pertinent and beneficial.

A huge text corpus is utilized to generate over 500,000 QAPs, which are subsequently used to train a QA system, demonstrating the approach's scalability. In addition, a web interface that users can engage with is offered, so they can create questions based on text input.

In conclusion, the study offers a novel method for producing high-quality QAPs through the identification of key phrases, the creation of questions, and the scoring of those questions with a deep learning-based model. The method exhibits encouraging results on a variety of datasets, and a large-scale QAP generation demonstrates its scalability.

The paper mostly focuses on creating questions, omitting the significance of creating excellent replies. A successful question-answering system requires both the creation of questions and the production of precise, illuminating answers. Future studies could look into ways to provide excellent responses to go along with the questions presented.

The report also lacks a thorough evaluation of the drawbacks and potential biases of the suggested strategy. It is still unknown whether the generated questions fully represent the variety of inquiries produced by humans. Additionally, the authors do not address any biases in the questions that are generated that may result from the training data. In order to solve these moral and social issues, more research is necessary.

# Chapter 3

# BanglaRQA Dataset

This section explains the whole data collection process of BanglaRQA, e.g., the source of the data, the criteria for data inclusion-exclusion, instructions given to the annotators, and other details. In addition, a comprehensive analysis of the dataset is included in this section.

## 3.1 Dataset Construction

The construction of BanglaRQA can be divided into 6 steps:

### 3.1.1 Passage Collection

The source for passages of BanglaRQA was Bangla Wikipedia. We collected 3000 passages manually covering a wide range of topics, from politics to sports, science to history, education to entertainment, and so on. We did the collection manually to ensure high-quality passages following the below-mentioned steps:

1. We chose passages focusing on a specific topic with no ambiguity skipping images, charts, and tables.

2. We either removed or translated, or converted words containing any other languages to Bangla.

3. We removed hyperlinks and citation numbers from the passages.

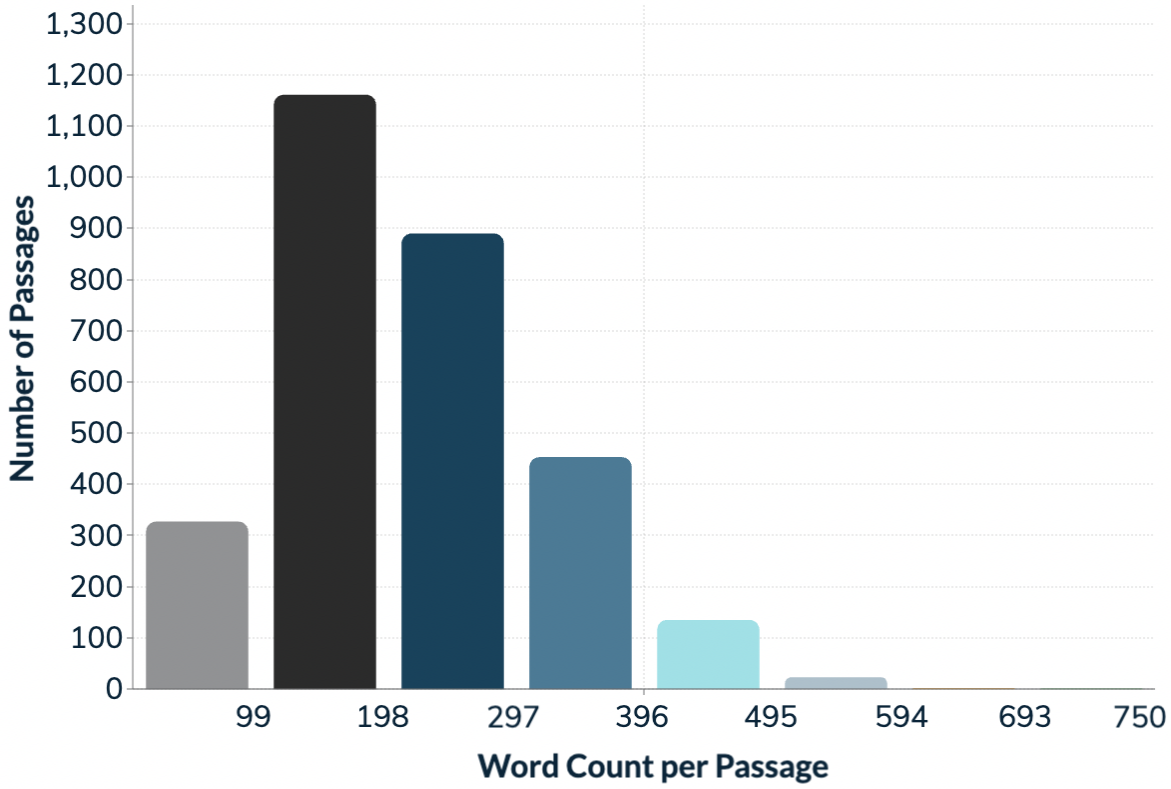


**Figure 3.1:** Distribution of word count per passage in BanglaRQA

Figure 3.1 shows the distribution of word count per context passage throughout the dataset. The average word count per passage is approximately 215 (1486 characters). To the best of our knowledge, passages in our dataset contain a higher character count than any previous Bangla reading comprehension datasets [25]. Hence, the Bangla language models will face the challenge of comprehending longer passages. After executing this procedure, we ended up with the passages setting up a good foundation for our dataset.

### 3.1.2 Crowd-workers' Recruitment

We recruited undergraduate engineering students from a prestigious institution by circulating a Google form explaining the purpose of the research and inviting them to apply. We then hired workers from applicants with at least 12 years of education in a Bangla-medium curriculum. Each annotator worked on 20 passages and was given a week to finish their work.

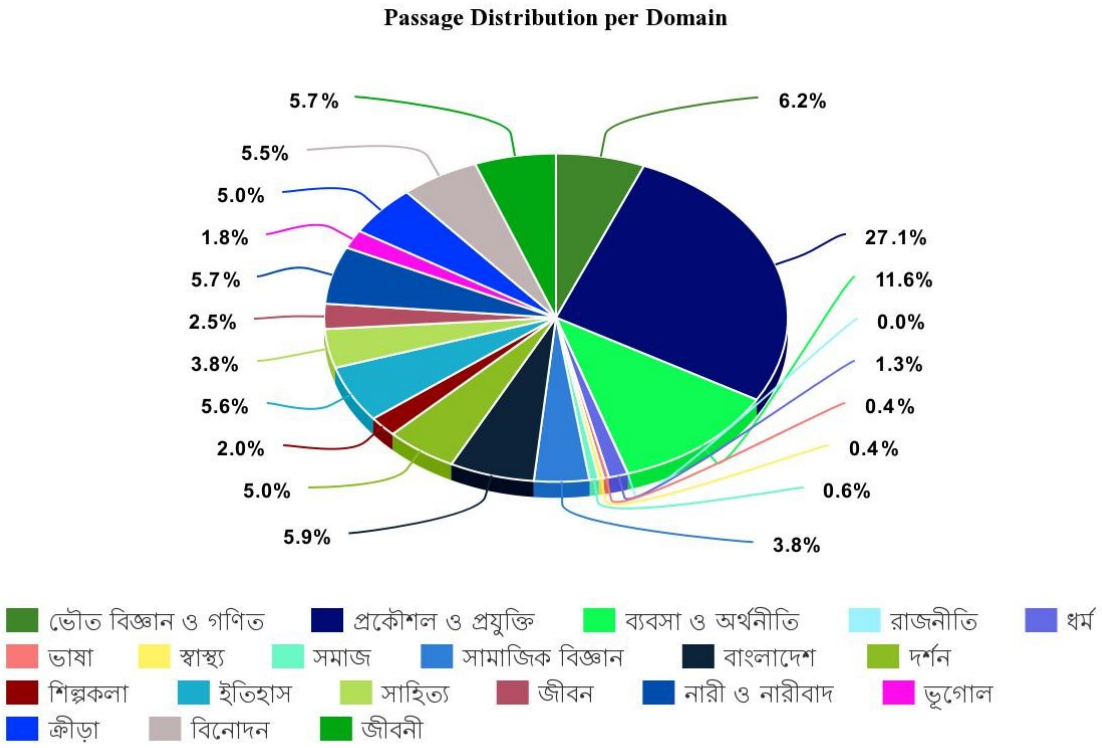


**Figure 3.2:** Passage distribution per domain

### 3.1.3 Question Collection

In this step, crowd-workers created questions from the previously gathered passages.
In order to make the questions lexically and syntactically as dissimilar to the context as possible, annotators were instructed to paraphrase and use synonyms as much as possible. From each passage, crowd-workers created 3 to 5 questions, out of which 1 or 2 questions were unanswerable. An unanswerable question means that the question cannot be answered using its corresponding passage. Crowd-workers marked each question as either answerable or unanswerable. There were no constraints given to them about the word limit per question. Figure 3.3 shows the distribution of word count per question. They also categorized each question into one of the following types:

- **Factoid Type:** This type of questions generally contain keywords like *What, Who, When, Where, Which* etc. Their answers are usually short phrases.

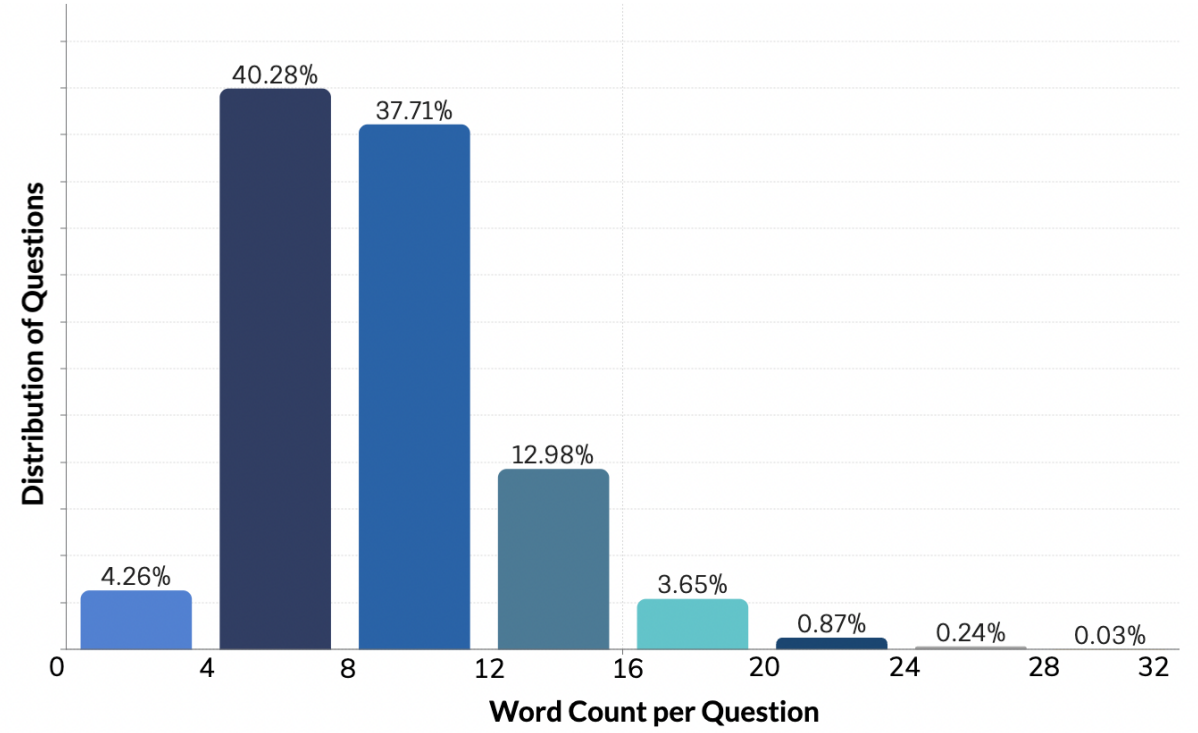


**Figure 3.3:** Distribution of word count per question in BanglaRQA

- **Causal Type:** This type of questions contain keywords like *Why, How* etc. Their answers are descriptive in general.
- **Confirmation Type:** This type of questions can be answered in *yes* or *no*. To answer confirmation-type question, often inference mechanism and higher level of knowledge is necessary.
- **List Type:** This type of question contains keywords like *What are..., Who are..., etc.* Their answers consist of multiple facts or entities.

Figure 1.1 contains an example question from each type with its respective passage and answer.

In total, we got 14,889 questions with a variety of different types combining both answerable and unanswerable ones from 3,000 passages.

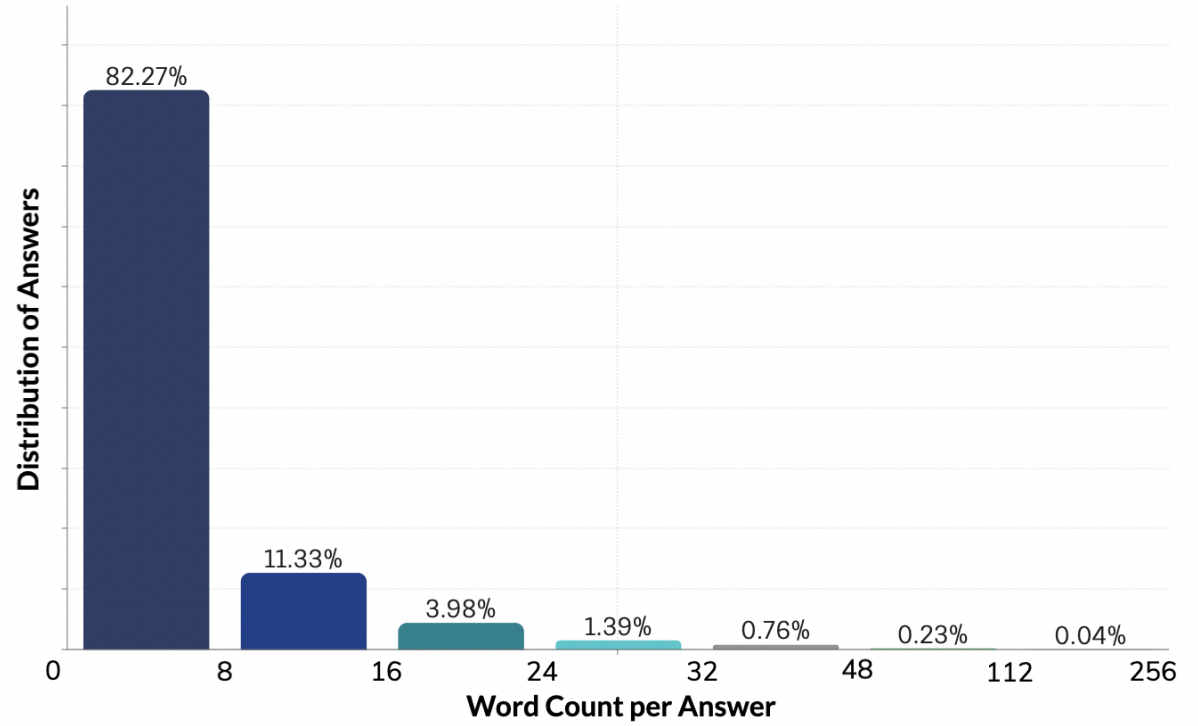


**Figure 3.4:** Distribution of word count per answer in BanglaRQA

### 3.1.4 Answer Collection

In this step, a different set of crowd-workers answered those questions from their corresponding passages. We gave them the passages with their respective questions. If they thought the question was answerable from the passage, they were asked to answer; otherwise, keep it blank. This was done to ensure the validity and quality of the questions.

Each question was answered by two different crowd-workers to increase the validity of the answers. Similar to question formulation, no word limit was given to the annotators as a constraint. Figure 3.4 shows the distribution of word count per answer. The crowd-workers then categorized each answer into one of the followings:

- **Single Span:** This type consists single shortest span from the passage correctly answering the question. These answers are primarily associated with Factoid type and Causal type questions.
- **Multiple Spans:** This type of answer consists of more than one span from different

parts of the passage separated by semi-colons (;). Factoid type and List type questions can produce this type of answers.

- **Yes / No:** Like the category name suggests, this type of answer consists of either *Yes* or *No*. Confirmation-type questions yield this type of answer.

Figure 1.1 contains an example answer from each type with its respective passage and question.

### 3.1.5 Quality Assurance

This step was crucial to ensure the overall quality and correctness of the dataset.
For the answerable questions, we checked if they were answered, and for the unanswerable questions, we checked if their answers were kept blank; otherwise, we marked those question-answer pairs as mismatches for later scrutiny.
Next, for each question, we checked the answers collected from two different annotators for their similarity and type. For list-type questions, we checked if the answers followed the fixed format where each entity was separated by a semi-colon (;). Again any mismatch was marked for later scrutiny.
The mismatched question-answer pairs were then given to a third set of annotators to validate. Their chosen response to the conflict was considered the appropriate one.
Finally, we compiled our dataset comprising 3,000 passages with 14,889 question-answer pairs where unanswerable questions had empty answers.

### 3.1.6 Train-Validation-Test Set Split

At first, we randomly split the passages into 80%, 10%, and 10% for our train, validation, and test set, respectively. Then, we sent each question-answer pair to the particular set where its associated passage belonged. So, our train set contains 2400 passages with 11,912 question-answer pairs, the validation set contains 300 passages with 1,484 question-answer pairs, and the test set contains 300 passages with 1,493 question-answer pairs.

## 3.2 Dataset Analysis

To better understand the contents of BanglaRQA, we analyze the distribution of different question-answer types for train, validation, and test sets. The split ratio of the question-answer pair for train, validation, and test set is approximately 8:1:1.

| Split | Unanswerable | Answerable |
|---|---|---|
| **Train** | | |
| Factoid | 2109 | 6220 |
| Causal | 409 | 730 |
| Confirmation | 218 | 1018 |
| List | 168 | 1040 |
| **Validation** | | |
| Factoid | 256 | 767 |
| Causal | 46 | 91 |
| Confirmation | 29 | 130 |
| List | 27 | 138 |
| **Test** | | |
| Factoid | 275 | 761 |
| Causal | 51 | 106 |
| Confirmation | 19 | 117 |
| List | 24 | 140 |
| **Total** | 3631 | 11258 |

**Table 3.1:** Dataset statistics of BanglaRQA based on question types

Table 3.1 shows the distribution of BanglaRQA based on question types in different splits. The ratio of answerable and unanswerable questions in each set is about 3:1. Under each split, the distribution for factoid, causal, confirmation, and list question types on unanswerable questions are around 72%, 15%, 7%, and 6%, respectively. For answerable questions, they are 68%, 9%, 11%, and 12%.

Table 3.2 shows the distribution for answer types of answerable questions. Each split has a percentage of single span, multiple spans, and yes/no at approximately 76%, 13%, and 11%. Here, one thing to note is the number of answers is less than the number of questions. This is because unanswerable questions doesn't have any corresponding answer parts.

The annotators had complete freedom to choose what type of questions they wanted to

ask, resulting in a higher percentage of Factoid questions. As the answers are dependent on questions, single-span answers followed the same trajectory as the Factoid questions.

| **Split** | **Single Span** | **Multiple Spans** | **Yes / No** |
|---|---|---|---|
| Train | 6835 | 1161 | 1012 |
| Validation | 850 | 148 | 128 |
| Test | 855 | 154 | 115 |

**Table 3.2:** Dataset statistics of BanglaRQA answerable questions' based on answer types

# Chapter 4

# Experimental Setup

## 4.1 Question-Answering System

The task is reading comprehension-based question-answering, where the model is given the question and its associated context passage as input. The model outputs answers in text format. If the question is unanswerable, then the output is an empty string. Four different models were implemented: BanglaT5, mT5, BanglaBERT, and mBERT. This section explains the whole pipeline of the experiments, from preprocessing the data to model training and evaluation.

### 4.1.1 Data Preprocessing

Questions, contexts, and answers all were first normalized [41]. Next, questions, context, and answers were all tokenized using the respective model's tokenizer. In the case of BanglaT5 and mT5, the maximum input and output lengths were 1024 tokens and 256 tokens, respectively. For BanglaBERT and mBERT, input and output were both 512 tokens. To ensure all the samples in a batch are of the same length, shorter inputs and outputs were padded, and longer ones were truncated.

### 4.1.2 Evaluation Metrics

As the task is reading comprehension-based question-answering, we used EM (Exact Match), and F1 score as a performance evaluation criteria. To calculate the F1 of multiple span type answers, we followed the DROP [42] paper. The other types of answers' F1 score calculation was similar to SQuAD 2.0 [38].

### 4.1.3 Models

| | **mBERT** | **BanglaBERT** | **mT5** | **BanglaT5** |
|---|---|---|---|---|
| **Optimizer** | Adam [43] | Adam | Adam | Adam |
| **No of epoch** | 15 | 15 | 15 | 15 |
| **Learning rate** | 2e-5 | 2e-5 | 5e-5 | 5e-5 |
| **Batch size** | 8 | 8 | 1 | 2 |
| **Time per epoch** | 7min | 7min | 75min | 48min |

**Table 4.1:** The training hyperparameters for different models

As BanglaRQA contains diverse answer types that were not available in any previously available Bangla datasets, namely, multiple spans and yes/no, it is necessary to establish baselines for both extractive and generative models. Therefore, we fine-tuned BanglaT5, mT5, BanglaBERT, and mBERT, state-of-the-art pre-trained Transformer models with different architectures for Bangla on the train set of BanglaRQA. Out of these, the first two models have the generative capability, whereas the other two have the extractive capability.
Multiple-span answers require information extraction from different parts of the passage. As a result, the standard question-answering BERT models that predict only the starting and ending token cannot provide multiple spans as answers. So, for extractive (e.g., BanglaBERT, mBERT) models, we followed the approach from [44] to accommodate multiple-span answers. Here, we considered it as a token classification task. For each token, the model predicts either 'B' denoting the start of an answer span or 'I' denoting other tokens of an answer span, 'O' if not part of any answer span. This approach can predict spans from different parts of the passage as an answer, which then can be merged to output the final answer.

The training hyperparameters are given in Table 4.1. We trained each of the models on the Google Colab Pro+ platform with P100 GPU. We saved the models after each epoch and calculated their EM and F1 score on the BanglaRQA's validation set. BanglaT5, mT5, BanglaBERT, and mBERT which performed the best on the validation set, were then evaluated on the test set of BanglaRQA. The test set results are given in Table 4.2.

| Model | EM | F1 |
|---|---|---|
| **mBERT** | 28.53 | 39.40 |
| **BanglaBERT** | 47.55 | 63.15 |
| **mT5** | 53.52 | 68.83 |
| **BanglaT5** | **62.42** | **78.11** |

**Table 4.2:** Performance of different finetuned models on BanglaRQA test set

BanglaRQA contains longer passages where information needs extraction from different parts of the passage. BanglaBERT and mBERT can only process 512 tokens at most, and the longer passages get truncated in the data processing step resulting in valuable information loss. Thus, their results are comparatively lower than BanglaT5 and mT5. Between the generative models, BanglaT5 performed significantly better as it was explicitly pre-trained for the Bangla language. Figure 4.1 shows the EM and F1 score on the validation set at each epoch for BanglaT5.

## 4.2 Question Generation System

To assess the generative capabilities of our system, we focus on developing a question-generation component. By providing the answer and context as input to BanglaT5 model, we aim to generate questions. The quality of the generated questions serves as an indication of the system's generative capabilities.

### 4.2.1 Evaluation Metrics

We used 2 evaluation metrics to measure the quality of our generated questions: BLEU Score and BERT Score. BLEU Score measures the similarity between the generated questions and

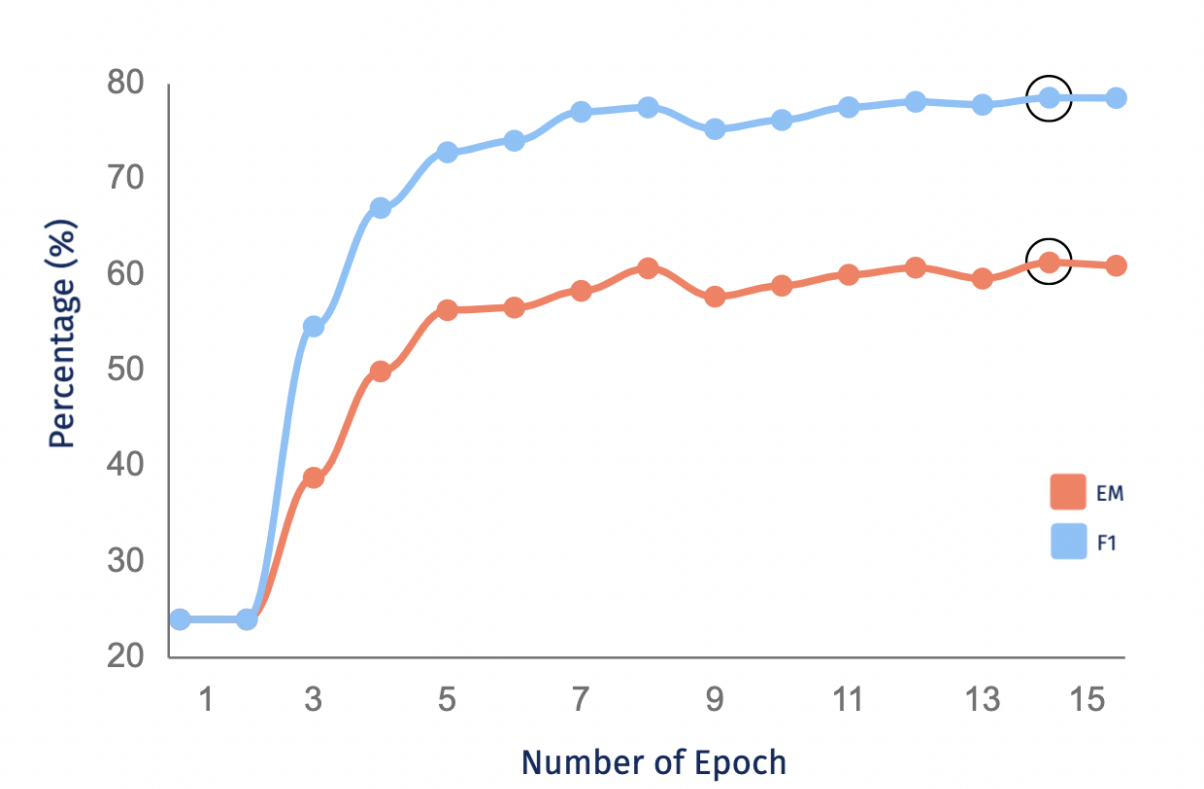


**Figure 4.1:** Performance of BanglaT5 on BanglaRQA's validation set

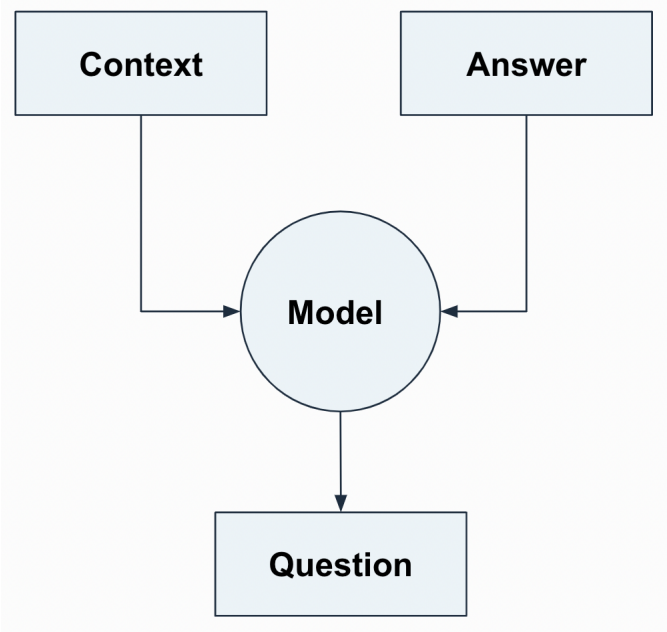


**Figure 4.2:** Question Generation Evaluation

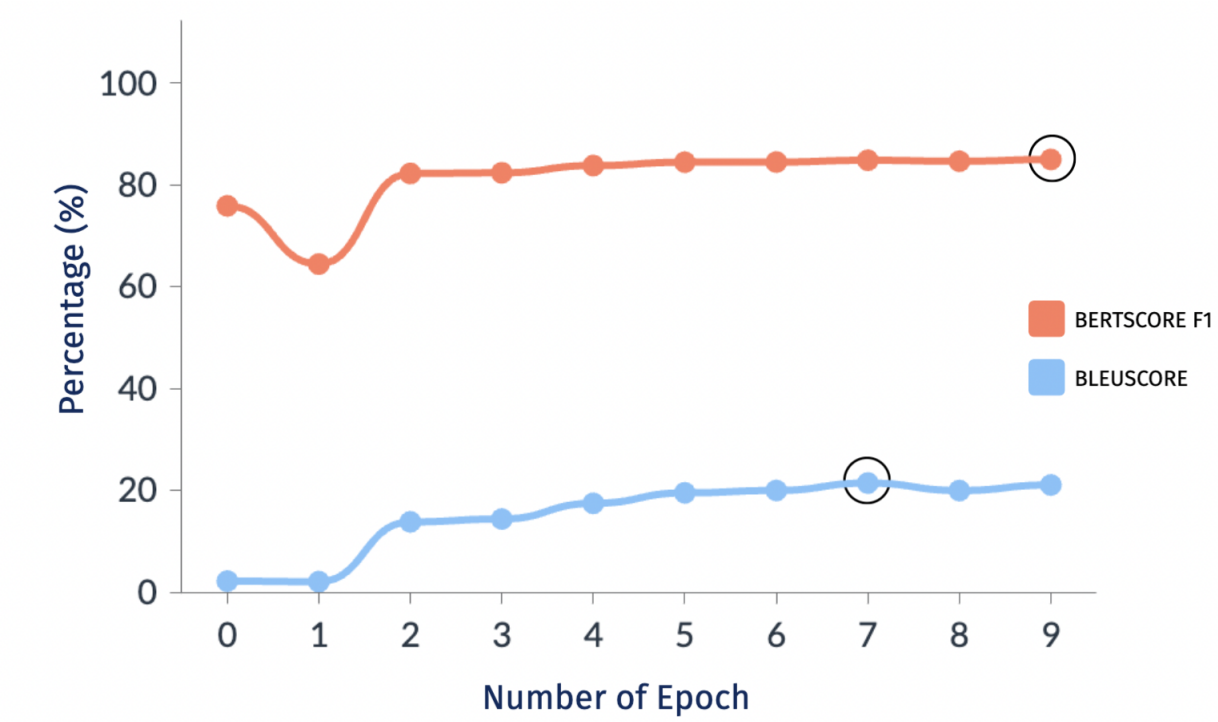


**Figure 4.3:** Evaluation on Model Generative Capability

the original questions in our dataset based on n-gram overlap, while BERT Score measures the semantic similarity between the generated questions and the original questions using a pre-trained BERT model.

### 4.2.2 Model

For this specific task, we employ the BanglaT5 model and utilize the Adam optimizer during the training process for a total of 9 epochs. Following the completion of the training phase, we analyze the outcomes and observe the following results:

| Evaluation Metrics | Score |
|---|---|
| **BLEU Score** | 21.56 |
| **BERT Score** | 85.04 |

**Table 4.3:** Performance of BanglaT5 on Question Generation

# Chapter 5

# Empirical Results and Analyses

After selecting the best model, BanglaT5, we evaluated the model on BanglaRQA's test set. It got an overall 62.42% EM and 78.11% F1 score. The upcoming subsections include analyses of model performance on different question-answer types and the utility of our dataset.

## 5.1 Performance on Answerable/Unanswerable Questions

The model got 54.89% EM and 75.73% F1 scores on the answerable questions, whereas it got 85.36% EM and 85.36% F1 scores on the unanswerable questions. As the model performance is lower for answerable questions, there is room for further improvement of the model.

## 5.2 Performance on Different Question Types

Table 5.1 contains model performances for all the different question types individually. The model performed best on confirmation-type questions as it just had to choose between two possible answers, *Yes* or *No*. The performance on factoid questions was a bit better than the overall performance as the answers for this type of question are mostly single-span. The EM and F1 score are comparatively low for the other two types of questions, especially for list type, as they require the accumulation of information from different parts of the passage.

| | Factoid | Causal | Confirmation | List |
|---|---|---|---|---|
| **EM** | 65.6 | 49.7 | 86.8 | 34.1 |
| **F1** | 80.3 | 69.4 | 86.8 | 65.2 |

**Table 5.1:** Performance of BanglaT5 on different question types

## 5.3 Performance on Different Answer Types

The model performance results on different answer types are available in table 5.2. The model performed with considerable accuracy on yes/no type answers. However, for multiple span type answers, the performance was the least, with only 20.78% EM and 55.75% F1 scores. This may be because the language model had to accumulate information from different parts of the passage.

| | Single Span | Multiple Spans | Yes / No |
|---|---|---|---|
| **EM** | 56.7 | 20.8 | 86.9 |
| **F1** | 77.8 | 55.7 | 86.9 |

**Table 5.2:** Performance of BanglaT5 on different answer types

## 5.4 Usefulness of BanglaRQA

To measure the usefulness of the BanglaRQA, we ran multiple experiments. We first trained the BanglaT5 model on our dataset and tested it on our test set. This provided us with 62.42% EM and 78.11% F1 score as shown in Table 5.3. From all the previously available Bangla question-answering datasets, only Bengali-SQuAD [22], and bn_squad are publicly accessible. Both are translated versions of SQuAD 2.0. However, bn_squad used a state-of-the-art translation process. Consequently, we compared BanglaRQA's generalizability to bn_squad. For that, earlier [27], when BanglaT5 was trained and tested on the bn_squad dataset (translated version of SQuAD 2.0) [23], it yielded 68.49% EM and 74.77% F1 score as shown in Table 5.3. Finally, we trained the model on our dataset and tested it on the bn_squad test set. This yielded 70.20% EM and 75.79% F1 score as shown in Table 5.3. This proves that

even though our dataset contains various answer types, it can successfully generalize on datasets like bn_squad with a single answer type (single span).

| Trained on | Tested on | EM / F1 |
|---|---|---|
| BanglaRQA | BanglaRQA | 62.42 / 78.11 |
| bn_squad | bn_squad | 68.49 / 74.77 |
| BanglaRQA | bn_squad | **70.20 / 75.79** |

**Table 5.3:** Performance of BanglaT5 for question-answering

## 5.5 Performance on Bangla Question-answer Pair Generation

We used BanglaT5 model to generate answer-agnostic Bangla question-answer pairs. We evaluated the quality of the generated QA pairs using two methods: human assessment and baseline comparison.

### 5.5.1 Human Assessment

For human evaluation, we randomly sampled a subset consisting of 115 generated QA pairs and manually checked them for 8 different pre-specified error types. The selected QA pairs were evaluated by three human evaluators who were proficient in the Bangla language and had a background in NLP. The evaluators were given instructions on how to assess the generated QA pairs for different error types. Each QA pair was evaluated independently by three evaluators and a majority vote was taken to determine the final label for the QA pair.

### 5.5.2 Baseline Comparison

Human assessment is not feasible for a large number of generated QA pairs. So, to have a holistic overview of the whole synthetic dataset we came up with the baseline comparison approach. We already established baseline performance for Question-Answering on the BanglaRQA dataset where fine-tuned BanglaT5 resulted in 62.42 EM and 78.11 F1 score.

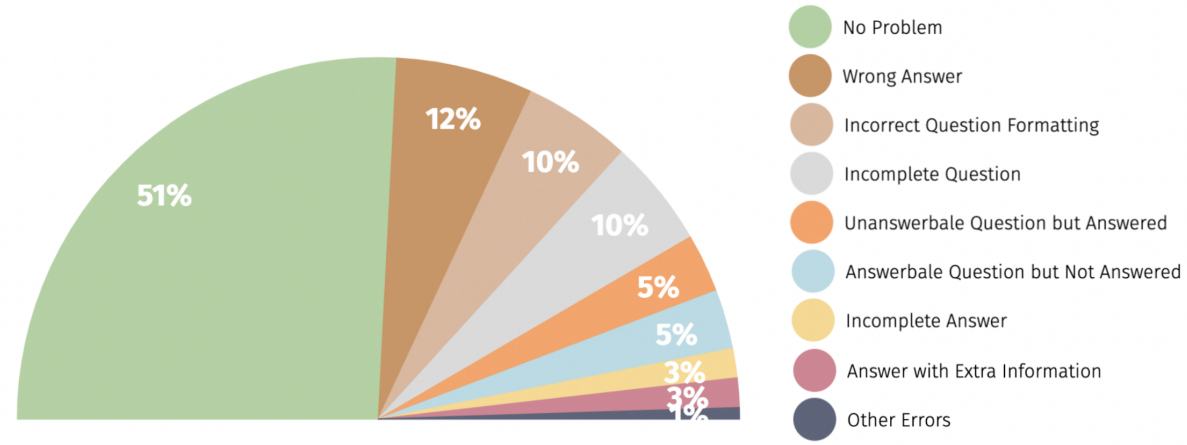


**Figure 5.1:** Human Assessment on Bangla Question Answer Pair Generation

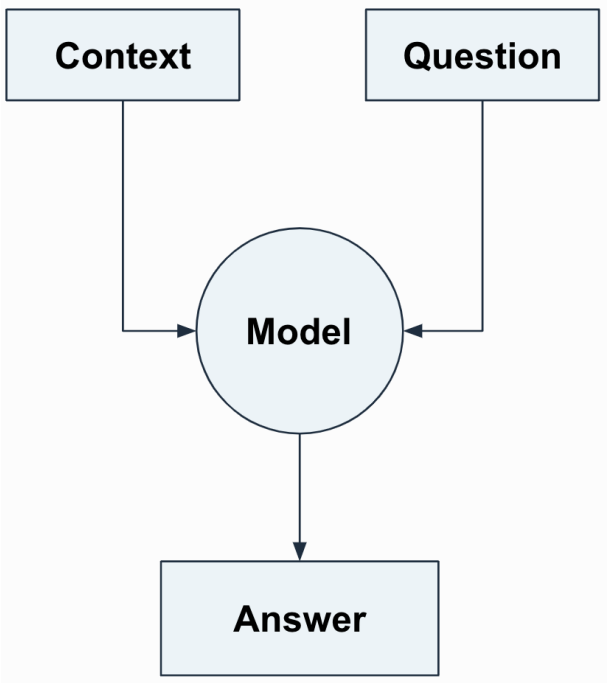


**Figure 5.2:** Baseline Comparison on Bangla Question Answer Pair Generation

For a baseline comparison, we combined the original BanglaRQA dataset and the generated QA pairs and trained BanglaT5 model on the augmented dataset. We then evaluated the performance of the fine-tuned model for Question-Answering (finding the answer to a question given a context) on this augmented dataset using the EM and F1 scores to compare with the baseline.

| Model | Dataset | EM | F1 |
|---|---|---|---|
| **Baseline (BanglaT5 for QA)** | BanglaRQA | 62.42 | 78.11 |
| **BanglaT5 for QA** | Augmented Dataset | 60.65 | 78.25 |

**Table 5.4:** Results of baseline comparison

The result indicates that the generated QA pairs show the same quality as human-annotated QA pairs.

# Chapter 6

# Conclusion

We introduce BanglaRQA, a benchmark dataset for Bangla reading comprehension-based question-answering with varied question-answer types. We finetuned both extractive and generative models to set baselines for our dataset. Upon training BanglaT5 on our dataset, we observed an overall performance of 62.42% EM and 78.11% F1 score.
However, the model could not do well on specific question types, e.g., list and causal, and specific answer types, e.g., multiple spans, which indicates some of the challenges of the dataset. Furthermore, testing our model on the bn_squad dataset yielded a better result, proving that our dataset is more generalized to bn_squad.
We then focus on Bangla Question-Answer pair generation. For Question generation, we use the answer-agnostic method. We show the generated question-answer pairs are of the same quality as human-annotated question-answer pairs through baseline comparison. In summary, this work presents a detailed end-to-end pipeline specifically designed for the task of Bangla question-answer pair generation.

## Limitations

We primarily encountered two challenges during the research process: human and computational resource. The limitation of human resources was a hindrance for us in creating a larger dataset. Due to constrained computing resources such as low-end GPU and limited memory, we could not pre-train our own language model. For this reason, we fine-tuned the existing

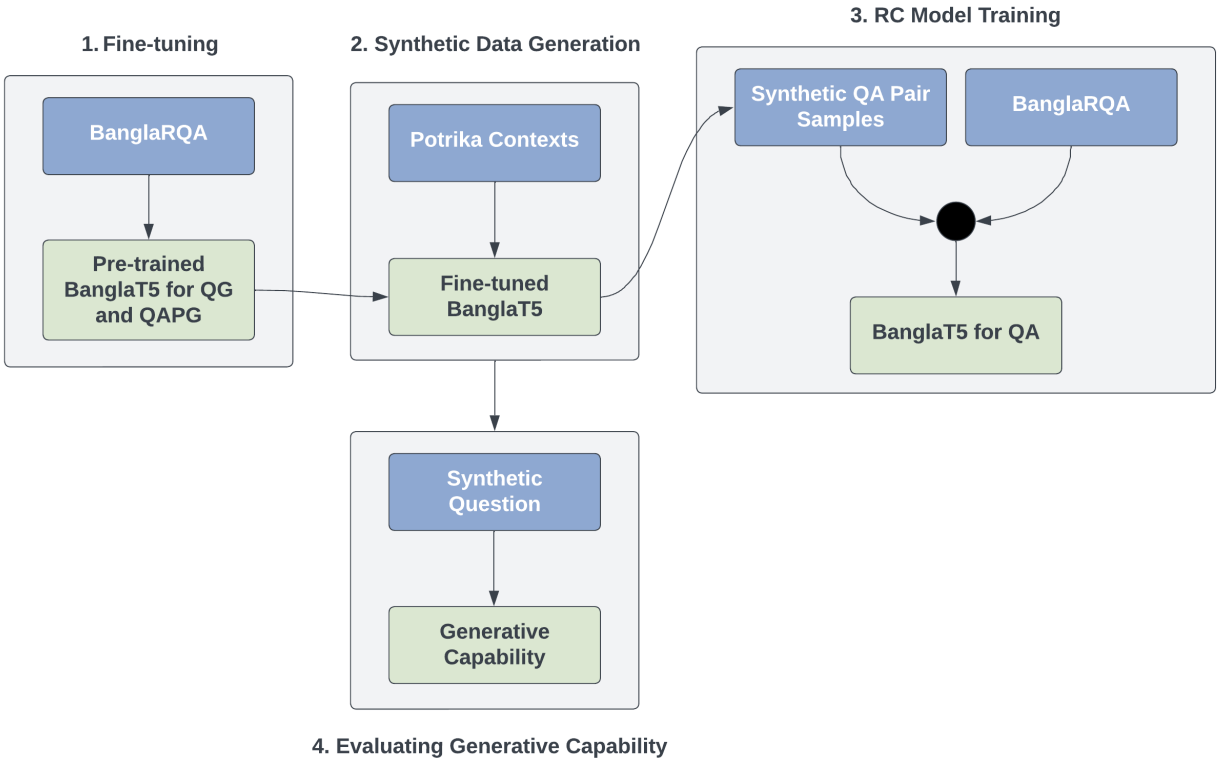


**Figure 6.1:** End-to-end Pipeline

pre-trained models. Hence, the performance was also dependent on their pre-training. Better pre-trained models may get improved accuracy on our dataset. All these limitations create future research scopes for Bangla reading comprehension-based question-answering.

## Future Scope

Our dataset can also be helpful in training language models for downstream tasks such as question-answering, answer-candidate generation, question generation, and question-answer generation. All of these have been shown useful in the English language to support other tasks, such as creating a Passage-QA index for retrievers, etc. Hence, we believe that BanglaRQA can be resourceful for further research on Bangla question-answering, question-answer pair generation and Bangla natural language understanding.

# Bibliography


[1] Huan Liu. Does questioning strategy facilitate second language (l2) reading comprehension? the effects of comprehension measures and insights from reader perception. *Journal of Research in Reading*, 44(2):339–359, 2021.

[2] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[3] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018.

[4] Simon Šuster and Walter Daelemans. CliCR: a dataset of clinical case reports for machine reading comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long Papers)*, pages 1551–1563, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[5] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[6] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. November 2016.

[7] Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Mo Yu, and Ying Xu. It is ai's turn to ask humans a question: Question-answer pair generation for children's story books. *arXiv preprint arXiv:2109.03423*, 2021.

[8] Sheng Shen, Yaliang Li, Nan Du, Xian Wu, Yusheng Xie, Shen Ge, Tao Yang, Kai Wang, Xingzheng Liang, and Wei Fan. On the generation of medical question-answer pairs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8822–8829, 2020.

[9] Vishwajeet Kumar, Kireeti Boorla, Yogesh Meena, Ganesh Ramakrishnan, and Yuan-Fang Li. Automating reading comprehension by generating question and answer pairs. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22*, pages 335–348. Springer, 2018.

[10] Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, pages 2032–2043, 2020.

[11] Sathish Reddy Indurthi, Dinesh Raghu, Mitesh M Khapra, and Sachindra Joshi. Generating natural language question-answer pairs from a knowledge graph using a rnn based question generation model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 376–385, 2017.

[12] Shivani G Aithal, Abishek B Rao, and Sanjay Singh. Automatic question-answer pairs generation and question similarity mechanism in question answering system. *Applied Intelligence*, pages 1–14, 2021.

[13] Michael Heilman and Noah A Smith. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617, 2010.

[14] Xuchen Yao, Gosse Bouma, and Yi Zhang. Semantics-based question generation and implementation. *Dialogue & Discourse*, 3(2):11–42, 2012.

[15] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*, 2017.

[16] Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv preprint arXiv:1603.06807*, 2016.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[19] Liam Dugan, Eleni Miltsakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, Dayheon Choi, Chuning Yuan, and Chris Callison-Burch. A feasibility study of answer-unaware question generation for education. *arXiv preprint arXiv:2203.08685*, 2022.

[20] Bingning Wang, Xiaochuan Wang, Ting Tao, Qi Zhang, and Jingfang Xu. Neural question generation with answer pivot. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9138–9145, 2020.

[21] Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th Interna-*

*tional Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE, 2021.

[22] Tasmiah Tahsin Mayeesha, Abdullah Md Sarwar, and Rashedur M Rahman. Deep learning based question answering system in bengali. *Journal of Information and Telecommunication*, 5(2):145–178, 2021.

[23] Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 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, Seattle, United States, July 2022. Association for Computational Linguistics.

[24] Mumenunnessa Keya, Abu Kaisar Mohammad Masum, Bhaskar Majumdar, Syed Akhter Hossain, and Sheikh Abujar. Bengali question answering system using seq2seq learning based on general knowledge dataset. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2020.

[25] Arnab Saha, Mirza Ifat Noor, Shahriar Fahim, Subrata Sarker, Faisal Badal, and Sajal Das. An approach to extractive bangla question answering based on bert-bangla and bquad. In *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, pages 1–6. IEEE, 2021.

[26] David Baidoo-Anu and Leticia Owusu Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Available at SSRN 4337484*, 2023.

[27] Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. Banglanlg: Benchmarks and resources for evaluating low-resource natural language generation in bangla. *arXiv preprint arXiv:2205.11081*, 2022.

[28] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 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 (NAACL): Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[30] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[31] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*, April 2020.

[32] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[33] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszko-

reit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

[34] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[35] Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

[36] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2174–2184, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[37] Md Asiful Haque, Shamima Sultana, Md Jayedul Islam, Md Ashraful Islam, and Jesan Ahammed Ovi. Factoid question answering over bangla comprehension. In *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–8. IEEE, 2020.

[38] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[39] Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 6027–6032, 2019.

[40] Husam Ali, Yllias Chali, and Sadid A Hasan. Automatic question generation from sentences. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 213–218, 2010.

[41] Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online, November 2020. Association for Computational Linguistics.

[42] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[43] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[44] Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3074–3080, Online, November 2020. Association for Computational Linguistics.