BanglaRQA: A Benchmark Dataset for Under-resourced Bangla Language Reading Comprehension-based Question Answering with Diverse Question-Answer Types

Syed Mohammed Sartaj Ekram, Adham Arik Rahman, Md. Sajid Altaf, Mohammed Saidul Islam, Mehrab Mustafy Rahman, Md Mezbaur Rahman, Md Azam Hossain, Abu Raihan Mostofa Kamal

Network and Data Analysis Group, Department of Computer Science and Engineering, Islamic University of Technology (IUT), Gazipur, Bangladesh

{sartajekram, adhamarik, sajidaltaf, saidulislam, mehrabmustafy, mezbaurrahman, azam, raihan.kamal}@iut-dhaka.edu

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, this paper introduces 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, this paper 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. Therefore, BanglaRQA has the potential to contribute to the advancement of future research by enhancing the capability of language models. The dataset and codes are available at https://github.com/sartajekram419/BanglaRQA

1 Introduction

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 (Liu, 2021). 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 (Rajpurkar et al., 2018), multiple documents as context (Welbl et al., 2018), fill in the blank type questions (Šuster and Daelemans, 2018), questions in natural language (Trischler et al., 2017), single span-based replies (Rajpurkar et al., 2018), answers in natural language (Nguyen et al., 2016), and so forth.

Bangla is the world's seventh most frequently spoken language, as over 230 million people speak it in Bangladesh and India (Karim et al., 2021). 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 (Tahsin Mayeesha et al., 2021; Bhattacharjee et al., 2022a), or very small datasets

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2518–2532
December 7-11, 2022

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, ব্রাউনিয় 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: 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. [1]

that focus on a specific topic area, such as general knowledge (Keya et al., 2020) or only for answer type (Saha et al., 2021). This illustrates the need for a diverse and high-quality dataset for NLP research in Bangla for question-answer based on reading comprehension.

To overcome these challenges, this paper presents BanglaRQA, a benchmark dataset for

[1]Samples with their English translation for each question-answer type are added on the *Appendix A.1* section

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: 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 (Bhattacharjee et al., 2022b), mT5 (Xue et al., 2021), BanglaBERT (Bhattacharjee et al., 2022a), mBERT (Devlin et al., 2019). 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 (detailed analysis is provided in section: 5), 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) (Bhattacharjee et al., 2022a), yielding 70.20% EM and 75.79% F1 score. Previously (Bhattacharjee et al., 2022b), 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.

Contributions. Our contributions are the following:

- 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.

2 Related Work

This section presents existing works for reading comprehension-based question-answering datasets in English and Bangla.

English Reading Comprehension Datasets

SQuAD 2.0 (Rajpurkar et al., 2018) and NewsQA (Trischler et al., 2017) 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 (Welbl et al., 2018), and HotpotQA (Yang et al., 2018) are datasets where the questions require finding and reasoning over multiple supporting documents to answer.

ReClor dataset (Yu et al., 2020) is collected from exams like GMAT and LSAT, making it very challenging. Another of this kind is RACE dataset (Lai et al., 2017), collected from middle school and high school English examinations in China.

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**	✔	✔	✔	✔

Table 1: Comparison among BanglaRQA and previously available Bangla reading comprehension datasets. Here, 'T' = Translation, 'HA' = Human Annotation, '✔' = Yes and 'X' = No

MS MARCO (Nguyen et al., 2016) 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 (Kwiatkowski et al., 2019) 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 (Khashabi et al., 2018) 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 (Reddy et al., 2019) 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 (Choi et al., 2018), where in a sample, a student asks a question about a hidden Wikipedia passage, and a teacher answers with spans from that passage.

Bangla Reading Comprehension Datasets

Some of the recent works (Tahsin Mayeesha et al., 2021; Bhattacharjee et al., 2022a) include creating translated Bangla datasets from SQuAD 2.0 (Rajpurkar et al., 2018). 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 (Keya et al., 2020) was also developed by using sources like Facebook and Google. Another dataset on factoid question answering (Haque et al., 2020) 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 (Saha et al., 2021) 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 (Rajpurkar et al., 2016) with no unanswerable questions and answers of only a single span from the context passage.

Table 1 illustrates a comparison among BanglaRQA and previously available Bangla question-answering datasets.

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:

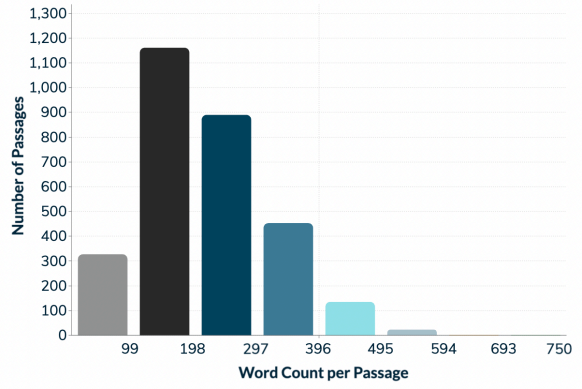

Figure 2: Distribution of word count per passage in BanglaRQA

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 (passage distribution given in *Appendix A.2*). 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 2 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 (Saha et al., 2021). 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.

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 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 কী (What), কে (Who), কখন (When), কোথায় (Where), কোনটি (Which) etc. Their answers are usually short phrases.
- **Causal Type:** This type of questions contain keywords like কেন (Why), কীভাবে (How) etc. Their answers are descriptive in general.
- **Confirmation Type:** This type of questions can be answered in হ্যাঁ (yes) or না (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 কি কি/ কোনগুলো (What are...),

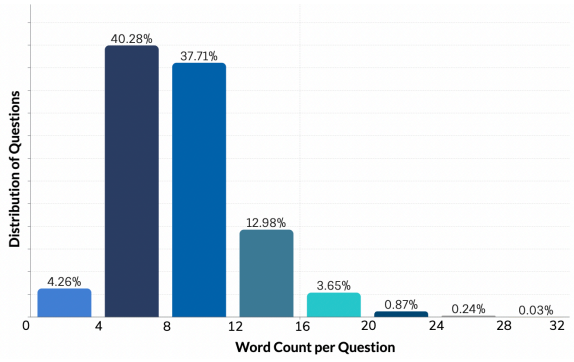

Figure 3: Distribution of word count per question in BanglaRQA

কারা কারা (Who are...), etc. Their answers consist of multiple facts or entities.

Figure 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.

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 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 হ্যাঁ (Yes) or না (No). Confirmation-type questions yield this type of answer.

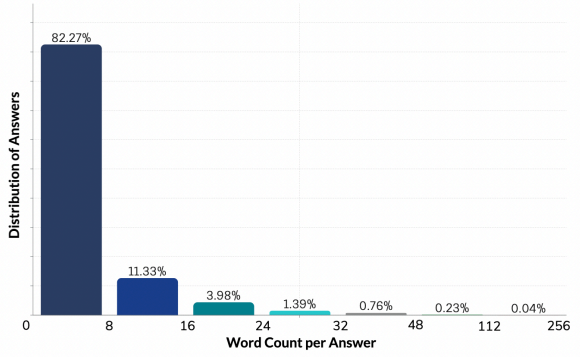

Figure 4: Distribution of word count per answer in BanglaRQA

Figure 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.

Table 2 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%.

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 2: Dataset statistics of BanglaRQA based on question types

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

Table 3: Dataset statistics of BanglaRQA answerable questions' based on answer types

Table 3 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%.

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.

4 Experimental Setup

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 Data Preprocessing

Questions, contexts, and answers all were first normalized (Hasan et al., 2020). 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.

4.2 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 (Dua et al., 2019) paper. The other types of answers' F1 score calculation was similar to SQuAD 2.0 (Rajpurkar et al., 2016).

4.3 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 (Segal et al., 2020) 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

	mBERT	BanglaBERT	mT5	BanglaT5
Optimizer	Adam (Kingma and Ba, 2014)	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: The training hyperparameters for different models

as an answer, which then can be merged to output the final answer.

The training hyperparameters are given in Table 4. 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 5.

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 5 shows the EM and F1 score on the validation set at each epoch for BanglaT5.

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

Table 5: Performance of different finetuned models on BanglaRQA test set

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.

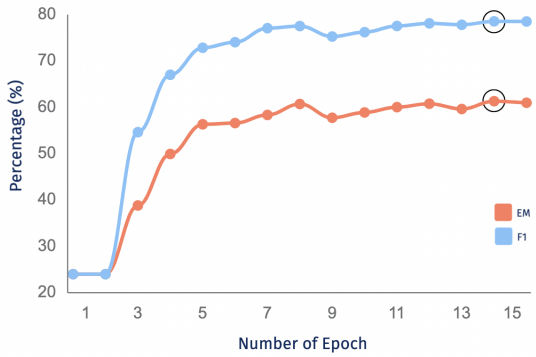

Figure 5: Performance of BanglaT5 on BanglaRQA's validation set

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 6 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, হ্যাঁ (Yes) or না (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.

5.3 Performance on Different Answer Types

The model performance results on different answer types are available in table 7. The model performed with considerable accuracy on yes/no type

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

Table 6: Performance of BanglaT5 on different question types

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.

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 8. From all the previously available Bangla question-answering datasets, only Bengali-SQuAD (Tahsin Mayeesha et al., 2021), 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 (Bhattacharjee et al., 2022b), when BanglaT5 was trained and tested on the bn_squad dataset (translated version of SQuAD 2.0) (Bhattacharjee et al., 2022a), it yielded 68.49% EM and 74.77% F1 score as shown in Table 8. 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 8. 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).

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

Table 7: Performance of BanglaT5 on different answer types

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. 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 and Bangla natural language understanding.

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 8: Performance of BanglaT5 for question-answering

Limitations

We primarily encountered two challenges during the research process: human and computational resource. Limitation of human resource 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 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.

Ethics Statement

We had to take a few things into account throughout the data collection and annotation process. We collected all our passages from Wikipedia-Bangla, an open-source, free encyclopedia in the Bangla

language. For data annotation, the annotators were provided adequate compensation (above minimum wages). Additionally, to protect the privacy of these annotators, all of their annotations were made anonymous.

Acknowledgements

This research is funded by IUT Research Seed Grant (RSG), reference number REASP/IUT-RSG/2022/OL/07/013. We would like to express our appreciation to our annotators for their assistance in the creation of the dataset, as well as our heartfelt thanks to Md Tahmid Rahman Laskar and Arowa Yasmeen for providing feedback. We also express our appreciation to anonymous reviewers for their comments and suggestions to enhance our paper.

References

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022a. 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. Association for Computational Linguistics.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. 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. Association for Computational Linguistics.

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

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. 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. Association for Computational Linguistics.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Mumenunnessa Keya, Abu Kaisar Mohammad Masum, Bhaskar Majumdar, Syed Akhter Hossain, and Sheikh Abujar. 2020. 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.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. 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. Association for Computational Linguistics.

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

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

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. 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. Association for Computational Linguistics.

Huan Liu. 2021. 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.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. 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. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. 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. Association for Computational Linguistics.

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

Arnab Saha, Mirza Ifat Noor, Shahriar Fahim, Subrata Sarker, Faisal Badal, and Sajal Das. 2021. 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.

Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. 2020. 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. Association for Computational Linguistics.

Simon Šuster and Walter Daelemans. 2018. 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. Association for Computational Linguistics.

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

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

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

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 (NAACL): Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. 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. Association for Computational Linguistics.

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

A Appendix

A.1 Samples of different question-answer types of BanglaRQA with their English translation

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 হোমিওপ্যাথিক চিকিৎসাকে মারাত্মক রোগের চিকিৎসা হিসাবে প্রচার করা ইত্যাদি।
অপবিজ্ঞানের বৈশিষ্ট্যগুলি হ'ল: অস্পষ্ট, অসঙ্গতিপূর্ণ, অতিরঞ্জিত বা অসমর্থিত দাবির ব্যবহার; নিশ্চিতকরণ পক্ষপাতিত্বের সাথে একটি কঠোর লড়াইয়ের দাবির প্রতিস্থাপন, বিষয় বিশেষজ্ঞদের দ্বারা যাচাই-বাছাইয়ের প্রতিরোধ করা; তত্ত্বটি বিকাশ করার সময় এবং পদ্ধতিগত পদ্ধতিগুলির অভাব। অপবিজ্ঞান শব্দটি অবমাননাকর হিসাবেও বিবেচিত হয়, কারণ এটি দাবী করে যে কোনও কিছুকে বিজ্ঞান হিসাবে ভুল উপস্থাপন বা ভুল উপস্থাপন করা হচ্ছে। সুতরাং, যাদের অপ বিজ্ঞানের প্রচার বা সমর্থক হিসাবে দেখানো হয় তারা এইটির বিরোধিতা করেন

Translated Context: Pseudoscience is a claim, belief or practice that is presented as science, but which does not follow the scientific method. If a research topic is presented according to the criteria of the scientific method, it does not follow these criteria. Malignancy can be harmful, for example, anti-vaccine workers present unscientific research that unjustifiably questions the safety of vaccines. Promoting homeopathic medicine as a treatment for serious diseases without any evidence etc.
The characteristics of science are: the use of vague, inconsistent, exaggerated or unsupported claims; The replacement of a rigorous fight claim with confirmation bias, preventing verification-selection by subject matter experts; Lack of time and methodological methods to develop the theory. The term malpractice is also considered derogatory, as it claims that something is being misrepresented or misrepresented as science. Thus, those who are portrayed as proponents or supporters of anti-science oppose it

Question: অপবিজ্ঞানের অপর নাম কী?
Translated Question: What is the other name of science?
Question Type: factoid
Is_Answerable: yes
Answer: সিউডোসায়েন্স
Translated Answer: Pseudoscience
Answer Type: single span

Figure 6: A sample (with English translation) of BanglaRQA with ***factoid*** question and ***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 কিছুই বুঝতে পারছিলেন না। এতে ক্ষুব্ধ হয়ে তিনি এটি অকিঞ্চিৎকর বই হিসেবে সরিয়ে রাখেন। কিন্তু পরবর্তীতে তার শিক্ষক আইজাক বারো তাকে বইটি আবার পড়তে বলেন। বইটি লেখা হয়েছিল দেকার্তের জ্যামিতিক গবেষণা ও কর্মের উপর।

Translated Context: He passed matriculation from Newton Trinity College in 181. While studying in college, he worked as a servant in different places of the college to cover the cost of his studies. There is no documentation from Trinity College that he did anything big as a student. However, it is known that he mainly studied mathematics and mechanics. At Trinity College, he first studied Kepler's theory of optics. He then turned his attention to Euclid's geometry. Because he could not understand some of the diagrams mentioned in a book of astrology bought from the fair. To understand these, Euclid needed to know geometry. Even so, Newton did not understand anything in the book. Angered by this, he removed it as a trivial book. But later his teacher Isaac Barrow asked him to read the book again. The book was written on Descartes' geometric research and work.

Question: কেন নিউটন কলেজের বিভিন্ন স্থানে ভৃত্যের কাজ করতেন?
Translated Question: Why did Newton work as a servant in different places of the college?
Question Type: causal
Is_Answerable: yes
Answer: পড়াশোনার খরচ চালানোর জন্য
Translated Answer: To cover the cost of tuition
Answer Type: single span

Figure 7: A sample (with English translation) of BanglaRQA with ***causal*** question and ***single span*** 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 অথবা তথ্য উন্মুক্ত নীতি নিয়ন্ত্রণ যেখানে নেই তা হতে পারে গোপনীয়তা সমস্যার প্রধান কারণ। অনেক উৎস থেকে ডাটার গোপনীয়তা সমস্যা তৈরি হতে পারে যেমন:
স্বাস্থ্য সচেতনতা রেকর্ড,
অপরাধী বিচারের তদন্ত এবং প্রক্রিয়া,
আর্থিক প্রতিষ্ঠান এবং বিনিময়,
আবাসিক এবং ভৌগোলিক রেকর্ড,
গোপনীয়তা ভেদ,
স্থানভিত্তিক সেবা এবং ভৌগোলিক অবস্থান।
ডাটা গোপনীয়তার প্রতিদ্বন্দ্বিতা হল ডাটা শেয়ারের সময় ব্যক্তিগত চিহ্নিতকরন তথ্য গোপন রাখার জায়গায়। ডাটা নিরাপত্তা এবং তথ্য নিরাপত্তা শাখা হার্ডওয়্যার, মানব সম্পদ ও সফটওয়্যারের নকশা এবং উপযোগিতার মাধ্যমে এই সমস্যা নিরসনে কাজ করে। যেহেতু ডাটা নিরাপত্তা সম্পর্কিত আইন এবং নীতিমালা প্রতিনিয়তই পরিবর্তিত হচ্ছে, তাই প্রতিনিয়ত আইনি পরিবর্তনগুলো গ্রহণ এবং নিরবিচ্ছিন্নভাবে ডাটা গোপনীয়তা এবং নিরাপত্তার নীতিমালার সাথে আপনার সম্মতি পুনমূল্যায়ন করুন।

Translated Context: Data Privacy or Data Privacy or Data Protection is the relationship between data, technology, public privacy expectations and the collection and distribution of legal and political matters.
Privacy relates to where personal identification information or other sensitive information is collected and stored (digitally or otherwise). Improper, ineffective, or lack of open policy controls can be a major cause of privacy issues. Data from many sources can cause privacy issues such as:
Health awareness records,
Investigation and process of criminal trial,
Financial institutions and exchanges,
Residential and geographical records,
Privacy breach,
Location services and geographical location.
The challenge of data privacy is to keep personal identification information secret while sharing data. The Data Security and Information Security Division works to address this issue through the design and utilization of hardware, human resources and software. As data security laws and policies are constantly changing, always accept legal changes and reassess your compliance with data privacy and security policies.

Question: তথ্য গোপনীয়তা কি ডিজিটালভাবে জমা হওয়া স্পর্শকাতর তথ্যের সাথে সম্পর্কিত?
Translated Question: Is data privacy related to sensitive information stored digitally?
Question Type: confirmation
Is_Answerable: yes
Answer: হ্যাঁ
Translated Answer: Yes
Answer Type: yes/no

Figure 8: A sample (with English translation) of BanglaRQA with ***confirmation*** question and ***yes/no*** 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। এটি উৎসর্গ করা হয় হিথ লেজারের স্মরণে, যিনি এর মুক্তির ছয় মাস আগে মারা যান। চলচ্চিত্রটি ২০০৮ সালের ১৬ জুলাই অস্ট্রেলিয়ায়, ১৮ জুলাই উত্তর আমেরিকায় ও ২৪ জুলাই যুক্তরাজ্যে মুক্তি পায়। চলচ্চিত্র সমালোচকদের মতে এই এটি সর্বকালের অন্যতম সেরা সুপারহিরো চলচ্চিত্র। বিশ্বব্যাপী চলচ্চিত্রটি ১ বিলিয়নের বেশি মার্কিন ডলার আয় করে। চলচ্চিত্রটি আটটি একাডেমি পুরস্কার মনোনয়ন পায় এবং সেরা শব্দ সংযোগের পুরস্কার জিতে। জোকার চরিত্রে অভিনয় করা হিথ লেজার সেরা পার্শ্ব অভিনেতার পুরস্কার পান। ২০১২ সালে এই চলচ্চিত্রের সিক্যুয়েল এবং সিরিজের শেষ চলচ্চিত্র দ্য ডার্ক নাইট রাইজেস মুক্তি পায়।

Translated Context: The Dark Knight is a 2006 American superhero film directed by Christopher Nolan. The film, based on DC Comics superhero Batman, is a sequel to the 2005 film Batman Begins. British actor Christian Bell played the role of Batman. Other cast members included Michael Kane, Heath Ledger, Gary Oldman, Aaron Eckhart, Maggie Jillenhall and Morgan Freeman.

The film is a joint production of the United States and the United Kingdom. It is dedicated to the memory of Heath Ledger, who died six months before its release. The film was released in Australia on July 17, 2008, in North America on July 18, and in the United Kingdom on July 24. According to film critics, this is one of the best superhero films of all time. The film grossed more than ১ 1 billion worldwide. The film received eight Academy Award nominations and won the Best Sound Connection award. Heath Ledger, who played the Joker, won the Best Supporting Actor award. The sequel to this film and the last film in the series, The Dark Knight Rises, was released in 2012.

Question: দ্য ডার্ক নাইটে কে কে অভিনয় করেন?
Translated Question: Who starred in The Dark Knight?
Question Type: list
Is Answerable: yes
Answer: ক্রিশ্চিয়ান বেল; মাইকেল কেইন; হিথ লেজার; গ্যারি ওল্ডম্যান; অ্যারন একহার্ট; ম্যাগি জিলেনহল; মরগান ফ্রিম্যান
Translated Answer: Christian Bell; Michael Kane; Heath Ledger; Gary Oldman; Aaron Eckhart; Maggie Jillenhall; Morgan Freeman
Answer Type: multiple spans

Figure 9: A sample (with English translation) of BanglaRQA with ***list*** question and ***multiple spans*** answer

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।

Translated Context: Einstein did a lot of research in various fields of physics and contributed a lot to new inventions and discoveries. The most famous contributions are the special theory of relativity (which integrated mechanics and electromagnetism) and the general theory of relativity (which established a new theory of gravitation by applying the theory of relativity to unequal motion). His other contributions include the cosmology of relativity, capillary action, chronological perceptual pigmentation, statistical mechanics and the solution of various problems of quantum theory which led him to explain the Brownian motion of the molecule, the probability of a molecular revolution, Religion (a theory of radiation that formed the basis of photon theory), gave the first idea of integrated field theory and geometricized physics.

Question: আপেক্ষিকতাভিত্তিক বিশ্বতত্ত্ব কী বোঝায়?
Translated Question: What does the cosmology of relativity mean?
Question Type: factoid
Is_Answerable: no
Answer:
Translated Answer:
Answer Type:

Figure 10: A sample (with English translation) of BanglaRQA with ***unanswerable factoid*** question

A.2 Passage distribution

Passages were taken from 20 different domains (approximately 130 sub-domains). The distribution of domains are as follows:

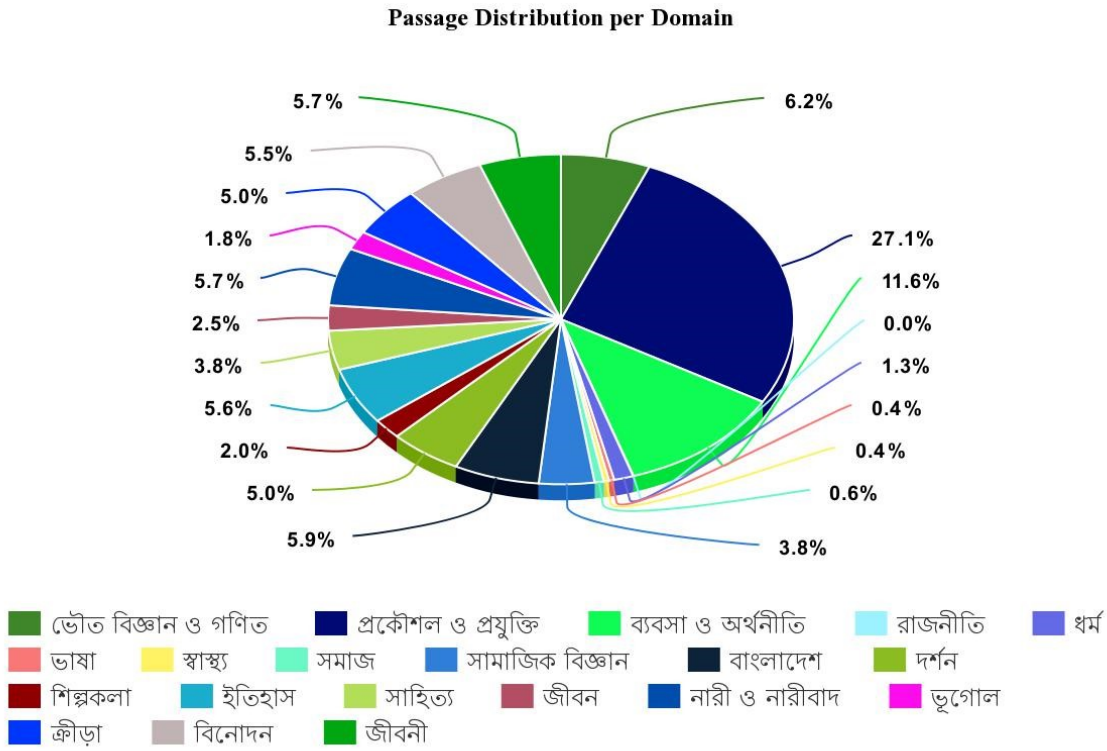

Figure 11: Passage distribution per domain