Reasoning Over Context in Bangla: A Generative QA Approach to Factoid Understanding Using LLM

By

Md. Masud Rana
Student-1 ID: 212-15-14760

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Mr. Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by

Mr. Abdullah Al-Amin
Lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

May 14, 2025

APPROVAL

This Project titled "**Reasoning Over Context in Bangla: A Generative QA Approach to Factoid Understanding Using LLM**," submitted by **Md. Masud Rana** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **14-05-2025**.

BOARD OF EXAMINERS

Dr. Fizar Ahmed
Board Chairman
Associate Professor, Department of CSE,
FSIT
Daffodil International University

Amatul Bushra Akhi
Internal Examiner 1
Associate Professor, Department of CSE,
FSIT
Daffodil International University

Mr. Abdullah Al Mamun
Internal Examiner 2
Sr. Lecture, Department of CSE, FSIT
Daffodil International University

Dr. Mohammed Nasir Uddin
External Examiner
Professor, Department of CSE,
Jagannath University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Mr. Md. Sazzadur Ahamed, Assistant Professor,** Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Mr. Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and Engineering Daffodil International University

Co-Supervised by:

Mr. Abdullah Al-Amin
Lecturer
Department of Computer Science and Engineering Daffodil International University

Submitted by:

Md. Masud Rana
Student ID: 212-15-14760
Department of Computer Science and Engineering Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Mr. Md. Sazzadur Ahamed, Assistant Professor** Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of my supervisor in the field of **Natural Language Processing (NLP)** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting themat all stages have made it possible to complete this project.

I would like to express my heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

I would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

This paper presents the development of a Bangla Question Answering (QA) system using advanced transformer-based models to tackle the complexities of Bangla language processing. Specifically, it compares the performance of BanglaT5, a model fine-tuned for Bangla, with mT5, a multilingual variant of the T5 model. Both models were evaluated on a dataset of over 7,500 Bangla news articles, focusing on factoid-based question answering. The results show that BanglaT5 outperforms mT5 on key metrics such as ROUGE, BLEU, Character Error Rate (CER), and Word Error Rate (WER), showcasing its superior ability to handle Bangla's unique linguistic features like morphology and syntax. BanglaT5 achieved a ROUGE-1 F1 score of 0.6979, Exact Match Accuracy of 0.49, and CER of 0.4054, demonstrating its ability to generate accurate, contextual answers. In contrast, mT5's performance was much lower, with an Exact Match Accuracy of 0.0008 and WER of 0.9996. This comparison highlights the importance of fine-tuning models for specific languages like Bangla, emphasizing the limitations of multilingual models in tasks requiring deep linguistic understanding. The system developed in this research offers a scalable solution for Bangla QA, with potential applications in education, public services, and digital literacy, contributing to the growing field of Bangla NLP. Future work will focus on deploying the model in real time, expanding the dataset, and exploring multimodal capabilities to increase its use in real-world applications.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This Chapter outlines the Introduction, Motivation, Objectives. Also, a short overview about the Methodology, Motivation and Project Outcomes also discussed.

1.1 Introduction

Natural Language Processing (NLP) A large source of progress in NLP has come from recent advances in using NLP to both understand and generate human language. Not all languages benefit from these advancements. High resource languages such as English, Chinese, and Spanish are supported by large standard dataset and computational resource, but low resource languages have to cope with these problems. Bangla is a major natural language and is spoken by about 230 million people worldwide but compared to world's total population, Bangla content on web are relatively less and the work on Bangla QA systems or Bangla NLP is still unappreciated. Efficient QA system has been made challenging largely due to the lack of the labeled data and the complexity of the grammar and syntax of Bangla [2].

Traditional rule-based QA systems and information retrieval methods have been already approached for dealing with QA tasks but these approaches are not able to handle the dynamic and complex characteristics. Especially for fields such as health care, education and general fact based questions. Recent works including BERT-Bangla have reported promising results for QA tasks in Bangla, though they still suffer from lack of improvement in closed domain references and complex domain centric enquiries [4].

This work attempts to alleviate those limitations by presenting a generative QA model for factoid questions in Bangla based on a fully fine-tuned large language model. In particular, we use the Bangla T5 model [5], which is adapted from text-to-text transformer architecture pre-trained on a large text corpus. We fine-tuned this model on a purpose-build dataset of 7,500 question-answers spanning different domains to make it capable of understanding and generating domain specific contextually relevant responses to factoid queries in Bangla.

The main aim of this research is to enhance the accurateness, fluency and coverage of Bangla QA systems. This work would contribute to low-resource NLP by building a scalable framework to extend a state-of-the-art transformer model such as T5 to less-resourced languages (or any low resource setting). The success in building this system will not only contribute to the advancement of Bangla NLP capabilities but also shows a

promising approach of employing AI-powered solutions to other low-resource languages even without extensive large-scale data.

1.2 Motivation

This research is motivated by the challenges in achieving modern Natural Language Processing (NLP) system's for Bangla Language. Despite being among the world's most widely spoken languages, it lacks the resources and data available for languages like English or Chinese Therefore, there remains a lack of complex system development like Question Answering (QA) system for Bangla. Current systems have, for the most part, focused on simpler tasks and far fewer have catered to more complicated problems such as fact-based QA. My primary contribution is to overcome these challenges by performing transfer learning on transformer-based models, such as the Bangla T5 model, for answering fact-based questions in Bangla. The work is to help AI models better understand and answer Bangla questions, so that the responses are more accurate and relevant. Solving this problem is beneficial for the progress of NLP in Bangla and AI systems for other low-resource languages.

Both researchers and Bengali speakers will benefit from this project. In doing so, it will also enhance AI tools for underrepresented languages and contribute to the creation of multilingual AI systems. Individually, it offers me an opportunity to contribute to a budding field and to expand access to information to millions of speakers of Bengali.

1.3 Objectives

The objective of this research is to develop an effective generative Question Answering (QA) system for the Bangla language using a fine-tuned Bangla T5 model. The specific objectives are:

1. To develop a generative QA system by full fine-tuning the Bangla T5 model on a dataset more than 7,500 real-looking factoid questions, passage and answers, ensuring reliable and context-sensitive response in Bengali.

2. To increase the accuracy and relevance of the system through improved training techniques, so that the model produces specific, context-aware answers in multiple fields.

3. To construct a flexible QA system that is suitable for questions in many domains, including history, geography and culture, and which can be applied to a wide range of different domains.

4. To enrich the area of Bengali NLP by extending the development for Artificial

Intelligence purposes for low resource languages and also to provide a platform which can be used by other uncollected languages.

5. Evaluating impact with the aim to enhance access to knowledge for Bengali speakers and increase use of AI tools.

1.4 Methodology

In this research, I take a systematic approach to develop a generative question answering (QA) system for Bangla by fine-tuning the Bangla T5 model. The process involves a number of crucial steps, including: data acquisition, models selection, model fine-tuning, and evaluation.

The process starts with the collection of a large dataset containing more than 7500 factoid questions and their answer (extracted from different domains, like history, political, literature, geography, culture and so on). These data were pre-processed with the upmost care to ensure healthy formatting, tailored for model input. The Bangla T5 model (a variant of the T5 architecture), well known for the capability of doing text-to-text learning model, was used as the backbone because of its versatility in complex text generation work. The model was further fine-tuned on our curated dataset, employing state-of-the-art deep learning methodologies to fine-tune the existing model on the particular characteristics of debsheds substantiation of answerable questions in Bengali Factoid QA. The fine-tuning focused on enhancing the accuracy, fluency, and context modeling for Bengali queries. To evaluate the model, the robust evaluation was performed with using standard evaluation metrics including accuracy, precision, recall, and F1 score. These metrics gave a global assessment of how good the model was for generating correct and sensible answers. Performance evaluations were conducted against prior models to demonstrate the developed system's improvement.

In final phase, we conducted comprehensive querying, robustness testing of the model, and spuriously correlation probing to check the generality of our model. Cross-validation methods were used to prevent overfitting and to ensure that the model performance was robust and generalizable.

1.5 Project Outcome

The expected contributions of this work are wide ranging, varying from theoretical results to application oriented studies. Key outcomes include:

1. Improved Bangla Question Answering System: The main achievements of this research will be the construction of a high quality generative QA model that can answer factoid type questions in Bangla with good precision and relevance. Not just improving the quality of Bangla NLP tools, this system will also add to the increasing number of researches that are focusing on AI systems for low

resource languages.

2. Benchmark for Future Research: This work will also serve as the benchmark for future research, in which fine-tuning the Bangla T5 model will be conducted and a well-annotated dataset with 7,500 factoid Q&As will be generated. This enriched corpus can be utilized as a benchmark for future work, for example, to facilitate the investigation of more complex issues, larger collections of data, and in different research domains, and so contribute to the development of AI for Bangla and other low-resourced languages.

3. Application: The developed system will facilitate better access to information for Bengali speakers, providing an AI enabled software which is expected to provide answers to natural factoid questions on multiaxial domains. The results have important implications for advancement of educational tools, automated helpdesk service and information-retrieval systems for Bengali, which will in turn encourage digital inclusion of Bengali speaking communities.

4. Contribution to Multilingual AI Systems: But, this study will focus on the complexity of low resource language (Bangla) and it will help us to develop the multi lingual AI system. The findings from this research could also be extended to other under-resourced languages and in general towards building more intelligent AI in the world-wild language technology.

5. Advanced AI Models applied to low resource languages: It will demonstrate the feasibility of the application of large transformer models in a resource-scarce landscape, and will show that high quality generative QA systems can be built even in low-resource scenarios. This result will encourage further investigation and investment on AI solutions for underrepresented languages in the wider scope of AI.

The proposed research will result in a powerful generative QA system, a step forward to improve the language processing capabilities of low-resource languages. It will also be an important platform for advancing multilingual AI in the future and contribute to digital inclusivity and to information access by Bangla speakers.

1.6 Organization of the Report

This report is structured into six chapters.

The chapter 1, presents the background formation for the study, motivation for study, objective, methodology, Project Outcome and Organization of the Report were all covered in chapter 1.

The Literature Review describes the related work done for Bangla NLP and QA systems,

Gap Analysis of previous work and necessity of this research summary is in chapter 2.

Chapter 3: Methodology – this section represents the methodology, including explanation about data collection, preprocessing techniques, model selection, and the fine-tuning procedure of our Bangla T5 model.

Chapter 4: Experimental Setup and Results – This chapter details the experimental setup of tools and environment and the results of testing the fine-tuned model against baseline models.

Chapter 5: Discussion and Implications – Here we put the findings in context, reflect on potential uses of the system as well as the social and practical implications, especially for Bengali speakers.

6.3 Conclusion and Future Work – This chapter provides the conclusion of the findings of the work, the contributions of our work, and discusses possible future work to refine and expand the syste

Chapter 2

Background

In this chapter, we present the necessary background and literature review related to Bangla QA task including the existing works that have been done and the methodology followed, issues addressed and open problems.

2.1 Introduction

This chapter provides the fundamental theoretical and contextual basis to understand the justification, extent, and technical direction of the presented work. The widespread availability of AI systems that can comprehend, interpret, and generate human language has dramatically influenced current human-computer interaction practices. In this respect, Question-Answering (QA) systems have played a relevant step forward, by permitting to the users a fine-grained and contextual relevant information retrieving from free natural language queries.

Despite being one of the most widely spoken languages worldwide, Bengali has been particularly under-resourced in terms of Natural Language Processing (NLP). "This has led to a significant traffic bottleneck for Bengali users, who are unable to access AI-based language technology and automatic information services. Though the state-of-theart models such as BERT and mT5 have revolutionized the performance benchmarks of multiple NLP tasks for high-resource languages, Bengali which is a low-resource language for QA is too naive for adaptation SJ Gupta et al: Preprint to Hasty Note) of these models and in the field of QA tasks adaptation of these two models is a recent endeavor and has not been much explored yet.

Thus this research is aimed at filling this technological gap by developing an effective transformer-based Bengali QA system. We have designed the system to understand user queries in the Bangla language and provide appropriate responses from fine-tuned pre-trained models. The background to this presented here positions the linguistic, technology based reasons behind these objectives and methodologies for this project.

In addition, the impact of QA systems in Bengali on society is very high. With growing concerns regarding the digital divide and equitable access, the need to provide intelligent language technology support to the Bengali speaking community is felt more urgently. Effective QA can contribute to educational development, administrative and governmental services, and user experience of digital systems. By breaking down the language barrier with which millions of native speakers are confronted, it not only champion's technological inclusivity, but also encourages the conservation and

modernization of the Bengali language in the digital sphere. This project is thus not simply a technical effort; it is a positive step toward linguistic freedom and the greater democratization of AI access.

2.2 Literature Review

The evolution of QA systems has experienced a huge change in the past few decades going to the initial rule-based models to modern deep learning and transformer based models. A Brief Survey of Question Answering Systems (Caballero, 2021) also described this journey from simple keyword retrieval systems like BASEBALL and LUNAR, to our current open-domain neural models that can comprehend complex natural language questions. The survey highlighted ongoing challenges, especially for multi-hop reasoning and the application of QA methods to low-resource languages like Bengali [14].

Hoque and Hasan (2019) were the frontiers of Bangla QA research with BFQA: A Bengali Factoid Question Answering System, and have used heuristics and statistical methods. Despite its performance, the system was somewhat restricted in providing answers to descriptive or multi-hop questions, even though it could answer 66% basic factoid questions correctly [4]. Similarly, Sarker et al. (2019) used statistical approach in their system which achieved an acceptable performance but was ineffective at handling queries of a complex structure [10].

The contribution was done by Das and Saha (2022), who used machine learning algorithms (ANN, SVM and Naïve Bayes) with the support of Word2Vec embeddings in their (Question Answering System Using Deep Learning in the Low Resource Language Bengali). Their question classification accuracy was as high as 95.88%, but answer extraction was difficult for complex queries [8].

In extension to the work of previous machine learning attempts, Khan et al. (2021) further progressed the area by proposing a model named Sequence-to-Sequence (Seq2Seq) that applied Long Short-Term Memory (LSTM) networks in their work Bengali Question Answering System with the Help of Seq2Seq Learning Based on General Knowledge Dataset. The proposed model attained an excellent training accuracy of 99 and validation accuracy of 89, which demonstrate that it can be powerful in learning a question-answer mappings from training data. However, authors noted that performance significantly deteriorated when the system processed, longer, syntactically complex, and semantically rich queries revealing the shortcomings of the Seq2Seq model to generalize across various linguistic constructions [3].

The introduction of transformer-based models was revolutionary step for Bengali QA. In their seminal work Unlocking the Potential of Multiple BERT Models for Bangla Question Answering in NCTB Textbooks, Khan et al. (2023) fine-tuned a number of pre-trained transformers (such as, RoBERTa and BanglaBERT) on educational datasets constructed

from National Curriculum and Textbook Board (NCTB) resources. They found that RoBERTa consistently produced better performance than other models according to accuracy, contextual understanding, and aptitude of handling various question structures [1].

Additionally, using transfer learning, Kabir et al. (2022) in which they investigated on multilingual BERT (mBERT) models fine-tuned with translated SQuAD 2.0 data. Their model achieved strong performance, and demonstrated that cross-lingual transfer learning was feasible for Bengali QA. However, significant issues were also identified: "adapting this model to the morphological variety and syntactic divergence of Bengali proved challenging, with precise answer retrieval often compromised" [2].

In an extension of this work in domain specific QA perspective Roy and Manik (2024) fine-tuned BanglaBERT to build a involving close domain QA system in terms of queries of Khulna University of Engineering & Technology (KUET). Our final model obtained an F1 score of 74.21% and an Exact Match (EM) score of 55.26%, which indicates that the model's good for processing institution-specific queries. However, the authors also observed that the size of the dataset, as well as the variety of domains, limited the generalizability more generally [11].

Considering the continued absence of large resources of high-quality questions that are necessary to develop a Bengali QA system, Ekram et al. (2022) presented BanglaRQA benchmark dataset which consists of 14,889 carefully created question-answer pairs. This dataset enabled standard evaluation procedures and provided researchers with a useful source for training and evaluating QA models. Models evaluated on this dataset performed impressively well with 78.11% F1 score, establishing a new benchmark for Bengali reading comprehension QA [9].

In the related area of Question Generation (QG), Ruma et al. (2023) proposed a better performing, answer-aware question generation model for Bengali language based on BanglaT5. We found that they achieved a BLEU-1 score of 38.57 and 98% grammatical accuracy, which shows that our model can generate contextually relevant and coherent and linguistically accurate questions. This study additionally reinforced the efficacy of the text-to-text transformer based models in the low-resource linguistic landscape of Bengali [5].

In order to investigate the wider scope of applications of LLMs in Bengali NLP, Maity et al. Patient/Provider Interaction Kazemian et al. (2024) did a substantive review of GPT-3. 5, GPT-4 Turbo and LLaMA models on grammatical error explanation. Their results show that this model GPT-4 is more robust than other models, having better explanations even for subtle grammatical structures. But, the research has found also the drawbacks for handling highly complex syntactic and semantic structures of Bengali language [6].

In the context of conversational QA for the institutions, Islam et al. (2024) developed PixieGPT, a modular, task oriented GPT based chatbot for university administration

tasks. The system was found to be domain-independent with strong domain-specific performance and capable of effectively understanding and responding to administrative questions. Although the model was highly successful in its own operational domain, it was less so domains outside that and more development is required [7].

In the very important area of health, Sen et al. (2024) introduced a Bengali healthcare question answering model by using state-of-the-art deep learning techniques. The system is able to answer medicine questions well with a filtered real-life medical interaction corpus. The authors reported that the initial results indicated a high level of accuracy and relevance; however, they highlighted the importance of rigorous validation and scalability tests to demonstrate the system's performance in practical clinical applications [12].

Lastly, the methodology for many of these progresses was established by Raffel et al. (2020) and his Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer paper which had presented the T5 model. Signature This is a great example of how a framework allowed for a quick fine-tuning of a pre-trained model on a specific problem, such as those posed in Bengali QA and QG research [13].

To conclude, Bengali QA research has advanced from primitive rule based and statistical techniques to state of the art transformer based models. There has also been remarkable progress in question type classification, context understanding and answer synthesis. In addition, the introduction of benchmark datasets and the usage of Large Language Model have constantly driven up the performance of the system. Nevertheless, data scarcity and the issue of complex and multi-hop queries are still challenges to be addressed along with standardized evaluation settings. These results demonstrate the progress made as well as the open research challenges to build strong and scalable Bengali QA systems.

Table 2.1 Summary of Literature Reviewed.

Author(s)	Year	Title	Methodology	Focus	Key Findings
Caballero[14]	2021	A Brief Survey of Question Answering Systems	Survey of QA history and methods	Broad QA system overview	Evolution from rule-based to neural models; challenges in low-resource languages
S. Banerjee et al.[4]	2019	BFQA: A Bengali Factoid Question Answering System	Rule-based and statistical QA	Bengali factoid QA	Effective for simple queries; not suitable for complex ones

Sarker et al.[10]	2019	Bengali Question Answering System for Factoid Questions	Statistical QA methods	Bengali factoid QA	Moderate accuracy; poor performance on complex queries
Das and Saha [8]	2022	QA System Using Deep Learning in Bengali	ANN, SVM, Naïve Bayes + Word2Vec	Question classification	95.88% classification accuracy; limited in answer extraction
M. Keya et al.[3]	2021	Bengali QA Using Seq2Seq Learning	LSTM-based Seq2Seq	General knowledge QA	99% training, 89% validation accuracy; degraded on complex queries
A. Khondoker et al.[1]	2023	Unlocking the Potential of Multiple BERT Models for Bangla QA	Fine-tuned RoBERTa, BanglaBERT	Education QA	RoBERTa achieved highest F1 score; superior contextual understanding
Das & Saha[2]	2021	Deep Learning Based QA in Bengali	Cross-lingual mBERT + SQuAD 2.0	Multilingual QA	Competitive EM & F1; domain adaptation challenges
Roy & Manik [11]	2024	Fine-tuning BERT-Bangla for Closed Domain QA	Fine-tuned BanglaBERT	University domain QA	F1 score 74.21%; limited by dataset size and domain
Ekram et al.[9]	2022	BanglaRQA	Benchmark dataset creation	Reading comprehension QA	Released dataset with 14,889 QA pairs; enabled robust evaluation
Ruma et al.[5]	2023	Transformer Based Answer-Aware Bengali QG	Fine-tuned BanglaT5	Question generation	BLEU-1: 38.57; 98% grammatical accuracy

Maity et al.[6]	2024	LLMs for Explaining Bengali Grammatical Errors	GPT-3.5, GPT-4, LLaMA	Grammatical error explanation	GPT-4 best performance; struggled with complex errors
Islam et al.[7]	2024	PixieGPT	GPT-based chatbot	University admin QA	Effective for administrative queries; closed domain only
Sen et al.[12]	2024	Healthcare QA in Bengali	Deep learning QA	Medical queries	Addressed healthcare queries; lacks large-scale validation
Raffel et al.[13]	2020	Exploring Limits of Transfer Learning (T5)	Pre-training and fine-tuning	Transfer learning, QA/QG	Text-to-text format yielded strong results across NLP tasks

2.2.1 Similar Application

Different kinds of research, technological approaches, and prototypes in relation to this work have also been investigated. Islam et al. (2024) presented PixieGPT, a modular GPT-style chatbot designed for university admin chat. Although the system was effective within its own context, it was not scalable and could not be easily adapted to other domains. Similarly, Sen et al. (2024) introduced a QA system tailored to the healthcare domain using deep learning. Even if the model showed ability to answer medical questions, its effective deployment is constrained by the scope and the breadth of its datasets and validations.

In the educational field, Khan et al. (2023) exploited fine-tuned BERT model in building a QA system on NCTB textbook texts and demonstrated the good performance of the transformer models in academic domains. They also fine-tuned BanglaBERT for closed domain QA at Khulna University of Engineering & Technology (KUET) generating impressive results but limited by small amount of data Roy and Manik (2024).

At the methodological front, Ruma et al. (2023) introduced an answer-aware Bengali Question Generation model, based on the down-stream task, demonstrated how answer-aware T5 lends itself to many tasks besides question answering, and revealed the newfound power of text-to-text transformers to generate question patterns with few errors in their grammaticality or the context with temperature=2.4. Cross-lingual methods as studied by Kabir et al. (2022) outperformed them by fine-tuning mBERT on translated SQuAD 2.0 and mBERT models, but it was harder for Bengali due to its morphological complexity.

Further, Maity et al. (2024) investigated LLMs (GPT-4 and LLaMA) as grammatical error explainer and their results are very promising, with limitations in managing syntactic

nuance. However, there are very few practical Bengali QA web or mobile applications available and most of them are experimental and claim reduced domain coverage and lack of linguistic diversity.

These collaboration efforts highlight the crucial need to build a scalable general-purpose Bengali QA system an aim that this research looks to accomplish by improving upon transformer-based approaches while working through the constraints identified with prior works

2.2.2 Related Research

The research in Bengali Question Answering (QA) systems have been seen to transition in a sequence from traditional rule based approaches, to statistical approaches, followed by to more modern deep learning and transformer based approaches. Earlier approaches, though well-suited for factoid questions, had no capability to handle complex, descriptive or multi-hop questions. Incorporation of classical machine learning algorithms such as ANN, SVM, Naïve Bayes along with Word2Vec embeddings further increased the accuracy for question classification, but still faced the challenge in context based answer retrieval.

The development of deep learning models, especially Seq2Seq architectures with LSTM, represented some progress, but there remained problems with syntax complexity. Transformer-based models, like BERT, RoBERTa, BanglaBERT, mBERT, etc., achieved even better performance by achieving significant improvements in term of situational awareness and the number of correct answers as demonstrated by Khan et al. (2023) and The difference is significantly observed between Roy and Manik (2024) . The development of benchmark datasets such as BanglaRQA standardised evaluation and paved way for further development.

There has also been a growing research interest in Question Generation (QG) and Authoring tools with Large Language Models (LLMs) such as GPT-3. 5, GPT-4, and LLaMA, with varied success in generative problems and grammatical error correction. Nevertheless, some issues are still not tackled, e.g., dataset diversity, cross-domain adaptation, and unified evaluation criterion. Between them, these studies demonstrate how far the field has come, and how much work there still is to be done work that this project will help support.

2.3 Gap Analysis

The gap analysis of studies on tea leaf disease detection highlights several key areas for improvement and future research. While many studies demonstrate significant strengths, such as high accuracy and computational efficiency, most are limited by narrow datasets, a lack of real-world validation, and restricted scalability. Datta and Gupta (2023) achieved robust classification accuracy of 96.56% and proposed IoT adaptability, but their dataset was limited to specific categories, necessitating the inclusion of additional diseases for broader applicability. Similarly, Gensheng et al. (2019) focused on computational efficiency but lacked testing on complex real-world scenarios, while Sun et al. (2018) excelled in preprocessing but did not explore transfer learning for scalability.

Krisnandi et al. (2019) used concatenated CNNs with moderate success, but their focus on only three diseases highlights the need for more extensive datasets and advanced architectures. Li et al. (2024) achieved high accuracy with MobileNetV3 and transfer learning but did not test across diverse datasets, limiting the generalizability of their findings. Latha et al. (2021) demonstrated computational simplicity in classifying eight diseases but did not validate their model in natural scenes or real-world environments.

Heng et al. (2024) introduced hybrid pooling for improved feature extraction but lagged in accuracy compared to other methods, emphasizing the need for dataset expansion and advanced optimization techniques. Bao et al. (2022) utilized multiscale feature fusion and attention mechanisms effectively but faced challenges in computational complexity and scalability testing. Soeb et al. (2023) achieved the highest accuracy (97.3%) with YOLOv7 for natural scenes but focused on only five diseases, suggesting the need for geographic and climatic diversity in datasets. Lastly, Rahman et al. (2024) provided a high-accuracy model for selected diseases in Bangladesh but lacked scalability for larger plantations or global contexts.

Overall, the studies showcase robust methodologies and promising results, but addressing gaps such as dataset diversity, real-world deployment, scalability, and computational efficiency is crucial for advancing tea leaf disease detection systems to practical, global applications.

Table 2.2 Summary of Gap Analysis.

Author(s)	Key Strength	Identified Gaps	Suggestions for Future Research
Khondoker et al. (2024)[1]	Effective use of multiple BERT models (RoBERTa, BanglaBERT)	Limited to educational domain; struggles with complex queries	Expand to open-domain QA; improve multi-hop reasoning

Das and Saha (2021, 2022)[8]	High question classification accuracy using ML & deep learning	Weak answer extraction; limited generalizability	Develop context-aware answer extraction; diversify datasets
Keya et al. (2020)[3]	Successful Seq2Seq model implementation	Poor handling of long/complex queries	Incorporate transformers or hybrid models
Banerjee et al. (2014)[4]	Early factoid QA prototype	Unable to handle descriptive or multi-hop questions	Apply modern deep learning models
Ruma et al. (2023)[5]	High grammatical accuracy in QG	Limited to QG; no QA capability	Extend model for full QA-QG integration
Maity et al. (2024)[6]	Effective grammatical error explanation using GPT-4	Focused only on GEE tasks	Apply LLMs to broader QA tasks
Islam et al. (2024)[7]	Domain-specific GPT chatbot (PixieGPT)	Limited to university admin queries	Generalize to open-domain and conversational QA
Ekram et al. (2022)[9]	Creation of BanglaRQA benchmark dataset	No support for conversational/multi-hop QA	Develop multi-hop and dialogue-based datasets
Sarker et al. (2019)[10]	Statistical QA approach	Limited accuracy; poor complex query handling	Transition to deep learning or transformer models
Roy and Manik (2024)[11]	Fine-tuned BanglaBERT for closed-domain QA	Dataset limitations; domain restriction	Use larger, multi-domain corpora
Sen (2024)[12]	Healthcare QA system proposal	Prototype stage; lacks scalability testing	Scale model and validate across medical domains
Das and Saha (2022)[8]	Effective cross-lingual transfer learning	Morphological and semantic challenges	Develop Bengali-specific pre-training techniques

Raffel and Shazeer (2020)[13]	Unified text-to-text transformer (T5) foundation	General framework; not specific to Bengali	Apply T5/mT5 to Bengali QA/QG tasks
Caballero (2022)[14]	Comprehensive QA system survey	Highlights research gaps but no implementation	Encourage empirical Bengali QA model development

2.4 Summary

In this chapter, we provided an overview of the background and related works for Bengali QA systems. 1 Introduction The importance of QA Technologies, especially in the domain of low resource languages such as Bengali was discussed in the introduction. The refiew of literature covered some of the existing works in rule-based and statistical methods, and in deep learning and transformer-based approaches with inspiration from previous related knowledge like BanglaBERT, RoBERTa and M5 models. Related applications in: education, healthcare, administrative domains, and methodological contributions, mainly on question generation and the use of Large Language Models (LLMs) were explored. In addition, our related work indicated an incremental improvement but still faced daunting challenges on dataset accessibility, domain transferability and handling for the complex query. Gap analysis showed a clear need for research and potential areas of breakthrough. Together, this chapter created the background and highlighted the need for building a state of the art Bengali QA system tasks we aimed to undertake in this project.

Chapter 3

Research Methodology

The chapter will provide a full discussion on design and research methodology of a context-aware Bangla Factoid Question Answering (FAQA) system. This involves modeling, data pick-up, model fine-tuning, performance evaluations, and system architecture. The motivation of selecting the transformer-based models (Bangla T5 and mT5) and the methodology of training and validating the model are explained. Moreover, detailed accessibility, scalability, and flexibility are maintained throughout the adopted methodological decisions.

3.1 Methodology

The project will leverage a state-of-the-art transformer-based large language model (LLM), such as Bangla T5 or mT5, fine-tuned using parameter-efficient adaptation techniques like LoRA to develop an advanced Bangla factoid question-answering system. The system will be designed to process a given question along with a relevant passage, then generate a precise and contextually accurate answer while ensuring efficiency and minimal computational overhead.

3.1.1 Overview

In this research employs transformer-based model for development of a context-aware factoid reading QA system for Bangla. The objective of the system should be to ingest user context and questions and utilizing it's high quality fine-tuned Large language model to produce and answer that will most closely fit the question. Two different models were used; Bangla T5, pre-trained on a number of Bengali language specific tasks and mT5, a multilingual model for comparative purpose. Our method is focused on transfer learning to discover how pre-trained linguistic knowledge on the source language can be adjusted for the Bengali factoid QA settings. The efficacy on models demonstrated that Bangla T5 was better in dealing with morphological and semantic challenges of Bengali in comparison with mT5. And that's how we could make Bangla QA efficient and scalable, compensating the limitation of language resources, and the QA model performance.

Furthermore, the model values robust semantic knowledge and flexibility, which enables the system to handle various question types and contexts adaptively. The work proposed here will have implications for the development of Bengali NLP and provide a base on which automated reasoning and question answering systems may in future be build.

3.1.2 Proposed Methodology

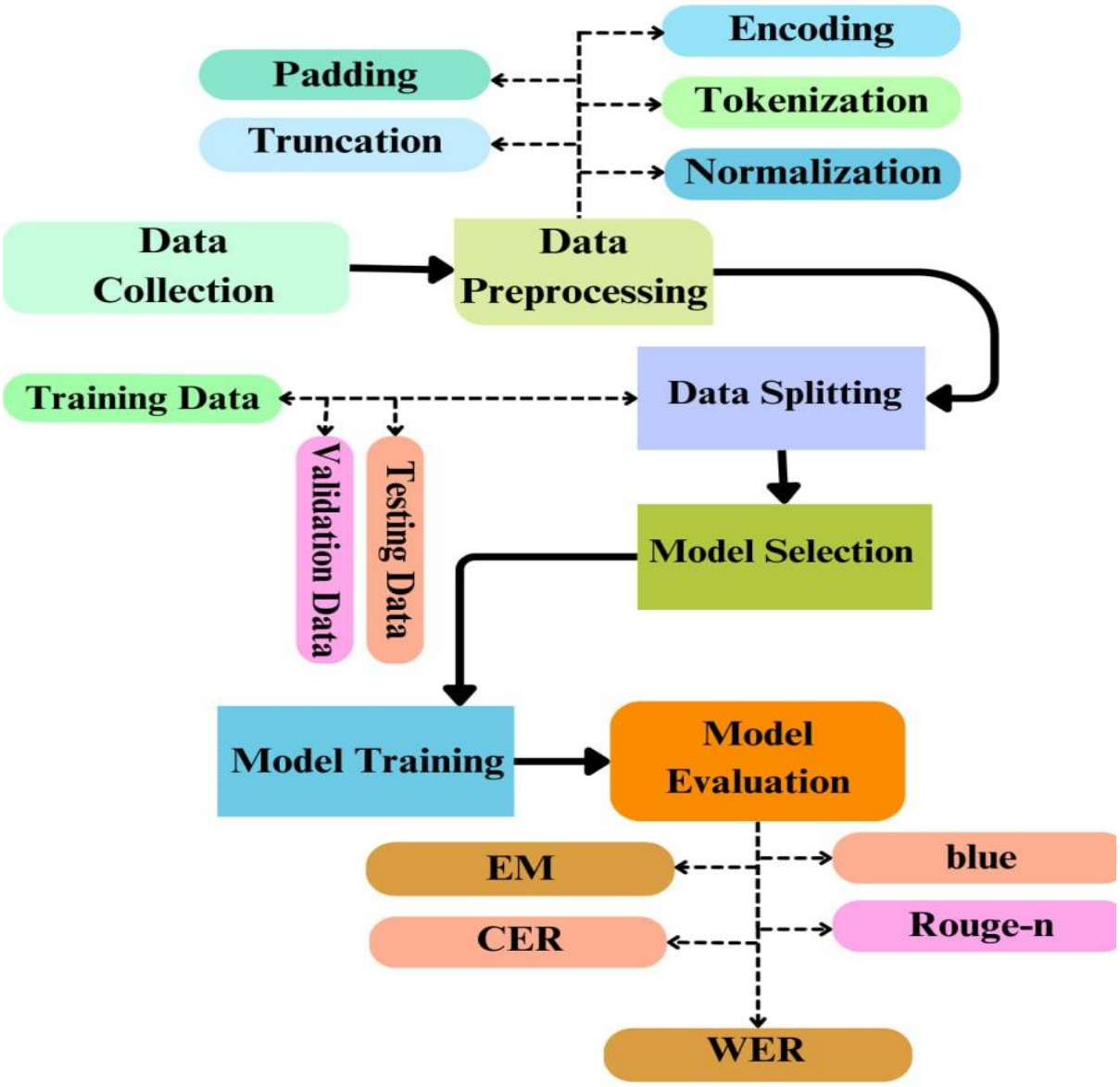

Figure 3.1 Proposed Methodology

The proposed methodology follows a structured pipeline designed to ensure efficient data processing, model training, and evaluation, as illustrated in Figure 3.1, It begins with data collection, where relevant data is gathered from diverse sources to ensure richness and variability. Once collected, the data undergoes a series of preprocessing steps to make it suitable for model ingestion. These preprocessing tasks include normalization to

standardize the text format, tokenization to convert sentences into tokens, encoding to represent these tokens numerically, and padding and truncation to ensure uniform input lengths across all sequences.

Following preprocessing, the data is split into three subsets: training data, validation data, and testing data. The training data is used to fit the model, while the validation data helps in tuning hyperparameters and preventing overfitting. The testing data is reserved for final evaluation to assess how well the model generalizes to unseen examples. After splitting the data, a suitable model is selected based on the task requirements and dataset characteristics. This selection may involve using existing architectures or fine-tuning pre-trained models.

The selected model is then trained using the training data. During training, the model parameters are optimized to minimize loss and improve performance on the validation set. This process continues until an optimal model is obtained. Finally, the trained model is evaluated using a comprehensive set of metrics. These include Exact Match (EM), which measures the percentage of correct answers that match exactly with the ground truth; Character Error Rate (CER) and Word Error Rate (WER), which quantify errors at the character and word levels, respectively; BLEU score, which evaluates the precision of n-grams between prediction and reference; and ROUGE-n score, which assesses recall-based content similarity. These evaluation metrics provide a detailed understanding of the model's accuracy, relevance, and fluency, thus validating its effectiveness in the target application.

3.2 Detailed Methodology and Design

The diagram outlines a structured methodology for detecting tea leaf diseases using machine learning techniques:

1. **Data Collection:**

   Data collection was very challenging part for me. I collected data more than 7500 from google search, comments on social media, Wikipedia, books, and other publicly available Bangla text sources and ensuring natural language presents. In Table 3.2 we can see a sample of my dataset. And mention the process are:

   Passage Selection: I have selected passage from various sources to maintain context as well as linguistic variety.

   Question: Then developed by hand Context-specific questions targeting linguistic differences and range of question difficulty.

   Answer: Responses were annotated manually to base upon the truthfulness and cohesion of the given text.

Review Dataset: The dataset was examined by two highly qualified academics in the fields of linguistics and NLP, who provided positive and constructive comments on the quality and relevance of the dataset.

Table 3.1 Dataset Sample

Question	Passage	Answer
বাংলাদেশের রাজধানীর নাম কী?	বাংলাদেশ একটি দক্ষিণ এশিয়ার দেশ। এর রাজধানী শহরের নাম ঢাকা, যা দেশের অর্থনৈতিক, রাজনৈতিক এবং সাংস্কৃতিক কেন্দ্র।	ঢাকা
বাংলাদেশে মোট কতগুলো জেলা রয়েছে?	বাংলাদেশে বর্তমানে ৬৪টি জেলা রয়েছে, যা দেশের প্রশাসনিক কাঠামোর গুরুত্বপূর্ণ অংশ।	৬৪
বাংলাদেশের জাতীয় সংসদ ভবনের স্থপতি কে ছিলেন?	বাংলাদেশের জাতীয় সংসদ ভবনের স্থপতি লুই আই কান।	লুই আই কান
পালরাজবংশের সবচেয়ে শক্তিশালী শাসক কে ছিলেন?	পালরাজবংশের সবচেয়ে শক্তিশালী শাসক ছিলেন ধর্মপাল। তিনি পাল সাম্রাজ্যকে একটি শক্তিশালী সাম্রাজ্যে পরিণত করেছিলেন।	ধর্মপাল
বাংলাদেশের সবচেয়ে বড় ব্রিজ কোনটি?	বাংলাদেশের সবচেয়ে বড় ব্রিজ হল 'পদ্মা সেতু', যা দেশের সবচেয়ে বড় এবং অত্যাধুনিক সেতু হিসেবে পরিচিত এবং দেশের যোগাযোগ ব্যবস্থার উন্নয়নে গুরুত্বপূর্ণ ভূমিকা পালন করছে।	পদ্মা সেতু

2. **Data Pre-processing**:

In this study, a number of preprocessing techniques were implemented in order to get the collected data ready for training the Bangla T5 and mT5 models, due to the linguistic neutrality and formatting of the data for machine learning purposes.

Cleaning text: We preprocessed the raw dataset to clean it by stripping unwanted punctuations, special characters and unwanted symbols to normalize the given text. This also required removing newlines and standardizing text formatting.

Tokenization: The tokenization is very vital step in converting our text data into format the model can understand. The AutoTokenizer was used to tokenize the input text (context+question) into its smallest processing unit called tokens. This was an effort to make sure all the text samples had been parsed correctly for analysis.

Stopword Elimination: Frequent Bengali stopwords (like "এটি", "এবং") were eliminated from the text to reduce the noise in the data. The removal of these non-content words caused the model to concentrate more on content words which help it to understand the context and answer the context adequately.

Normalization of texts: All texts were normalized in order to mitigate the differences in writing scripts, spelling and format. For example, this involved making the script uniform and lower casing all the text, so all the tracks are consistent across the dataset.

Padding and Truncation: The sequences that are shorter than the model's maximum input length are padded with zeros and the ones that are longer are truncated to the maximum length. For Bangla T5 and mT5, we padded/truncated the input sequence to a maximum of 1024 tokens and the output sequence to a maximum of 256 tokens. This ensured the model always got equal-size segments, which was important for the training quality.

Data Split: Following pre-processing, the data were split into training, validation, and test sets. This partitioning made it possible to train and test the model on unseen part of the data avoiding both model overfitting and bias.

It was critical have these preprocessing steps in place to have a clean and well-structured data to facilitate fast model training, which in-turn enabled Bangla T5/mT5 models to do well over the Bangla Factoid QA tasks.

3. Train Model:

The presence of Bangla T5 and mT5 is crucial to address the specific tasks of Bengali Question Answering (QA). Bangla T5, specially pre-trained for Bengali, is capable of capturing the complexity of the language, specifically its rich morphology, complex syntax and diverse semantics, in its highly accurate and contextually relevant replies. On the other hand, the state-of-the-art mT5 -based multilingual model allows us a comparison and the success of a cross-lingual model for Bengali QA tasks is converged here. This twin-model method will give a good trade-off between a language-specific model and a multilingual model. Architectures such as BERT perform well for extractive QA, but are unsuitable for the requirements of this project because they are not generative. In contrast to BERT that read a fixed passage for generating answers, Bangla T5 and mT5 produce answers from context and thus are more suitable for generative QA systems. Collectively, the models present a fine balance between specialization and generality, providing a scalable framework for future developments while establishing a strong baseline for advanced Bengali NLP tasks. Both model follows encoder and decoder architecture, show in Figure 3.2.

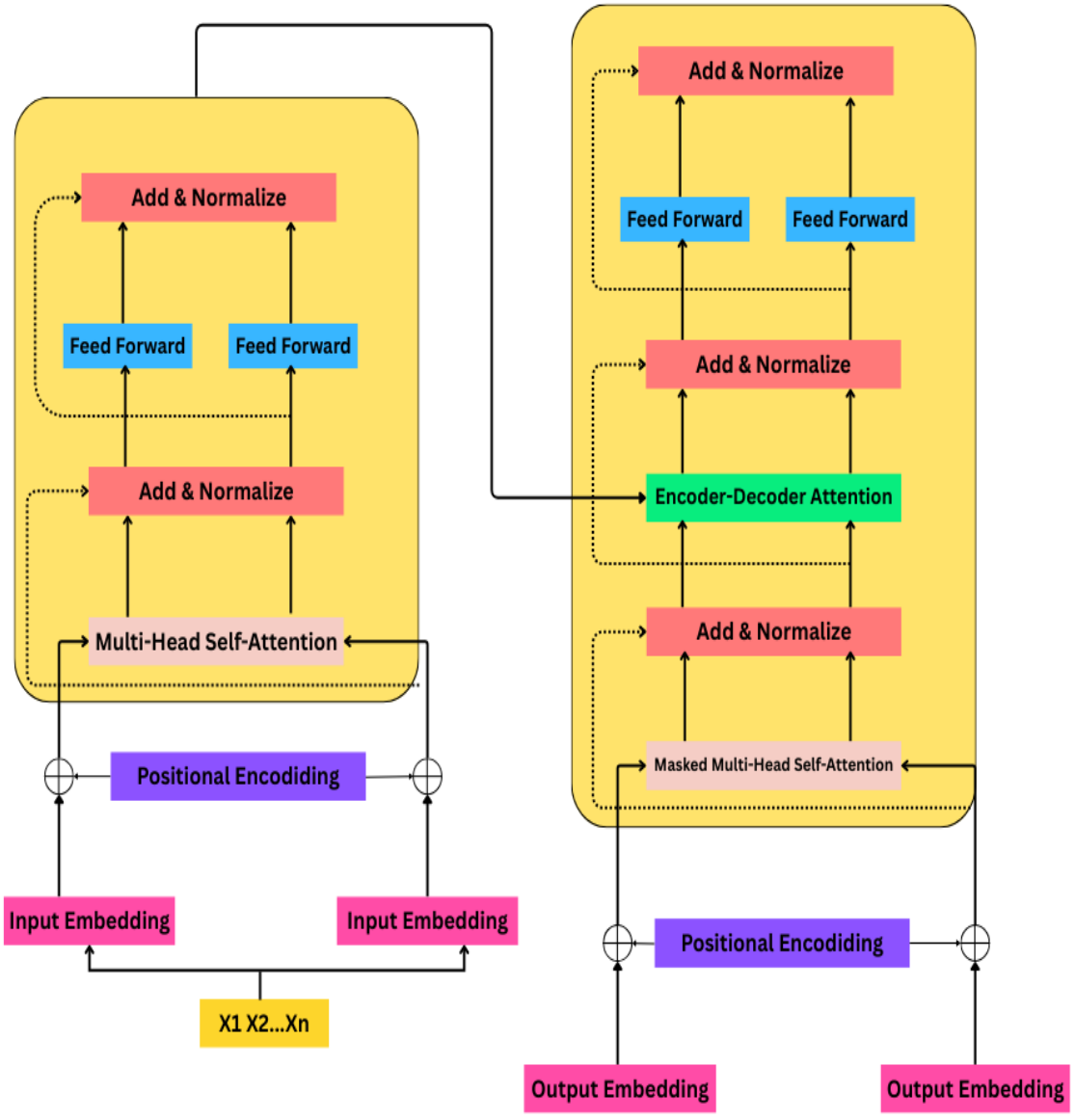

Figure 3.2 Encoder-Decoder Transformer Architecture

Bangla T5:

Bangla T5 is a fine-tuned Bengali version of the T5 (Text-to-Text Transfer Transformer) model and performs very well on Bengali NLP tasks. It is built on the T5 architecture, where all NLP tasks are formulated as a text-to-text problem. The model was specially fine-tuned with Bengali language data to improve its performance on tasks like question answering, text summarization, text classification. Bangla T5 has some ~220M parameters in base to allow it for space and time efficient execution on Bengali task. The model adopts Sentenc Piece tokenization, which is well-suited for the script and morphology of Bengali. BT5 performed well on ROUGE, BLEU, F1 scores, particularly in summarization and factoid question answering tasks. Its performance on Bengali text

generation, summarization and other NLP tasks demonstrate that the model's fine-tuning output results in contextually appropriate and coherent answers.

Key Components:

Encoder: The encoder of the Bangla T5 receives input text that is formed by context (i.e the passage) and the question. It tokenizes the text, and applies multi-head self-attention and feed-forward neural networks. This permits the model to learn contextual information about the words in the passage and the question. The output of the encoder is a contextually rich representation of the input. Positional Encoding preserves the order of tokens in a sequence, which is needed for deciphering sentence structure and meaning.

Decoder: The decoder constructs the output according to the context representation provided by the encoder. It leverages cross-attention to attend relevant portions of the context on generating the response. The decoder is set up to produce a series of text, which is great if you need tasks where the model needs to construct answers from context rather than simply extracting them. The Masked Multi-Head Attention in the decoder ensures that the model only looks at the attention words in the proper order of generating and never in the future.

Self-Attention Mechanisms: Self-attention mechanism enables the model to attend to different parts of the sequence (both the context and the question) in order to capture word-word relationships at various positions. This is crucial for modelling long-range dependencies and context in natural language.

Feed-Forward Networks:In addition to the attention layers, it is followed by a subsequent multi-layer of fully connected feed forward networks to refine the context representation and help the model in generating the most likely answer.

Output Generation: The decoder generates the resulting output, which in a generative QA system such as Bangla T5, is a well-formed, grammatical Bengali answer. The output is decoded to a natural text for the purpose of user display.

Advantages of Bangla T5 Model:

Generative Capacity: In contrast with extractive models, which only pick an answer from the context, Bangla T5 generates an answer according to its understanding of the passage and the question. This makes it fitting for tasks where answers are not explicitly mentioned in the passage but involve reasoning over the text.

Text-to-Text Architecture: The T5 architecture is able to handle all NLP tasks as text generation task and hence is a single architecture that can be used for several tasks like QA, summarization, translation etc. without the need for different architectures.

Can be Adapted to Bengali: Bangla T5 is finetuned on Bengali dataset which makes it capable to process the peculiarities of the syntactic and morphological structures of Bengali language to achieve better performance than generic models.

In conculation, The Bangla T5 model, an encoder-decoder style model possesses the generative properties that are best fit to solve a Bengali factoid question answering task. It learns the input context and produces context aware answers making it a strong candidate for any QA system in Bengali.

Multilingual T5:

mT5 (Multilingual T5) is a multilingual variant of the T5 model, with training examples in 101 languages based on mC4 (multilingual C4) which is a multilingual version of the C4 dataset. It is developed to be cross-linguistic and encompasses a vast array of languages. mT5 comes in some sizes -mT5-small (300 million parameters), mT5-base (580 million parameters) and mT5-large (1.2 billion parameters). Although mT5 is highly multilingual, it is not optimized for Bengali, and might not be as effective as BT5 in tasks that demand a deeper understanding in Bengali. It employs the model's default SentencePiece tokenization which although has been shown to effectively model multiple languages, may fail to capture the complete richness of the Bengali language including its highly agglutinative morphology and graphological subtleties. So as mT5 scores well on a wide range of languages, BT5 gives better localization when it is used in Bengali-specific NLP tasks such as generating Bengali text, question-answering over Bengali data, and Bengali summarization. However, it provides a benchmark model and comparative results of specialized versus multilingual models for the task of Bengali QA.

Key Components:

Encoder: The encoder takes the input text, which could be in any of the supported languages. It leverages multi-head self-attention and feed-forward layers to capture the word relations within the context.

Decoder: Given a description of the encoder's input representation, the decoder produces output (answers) and cross-attends to the most relevant parts of the context.

Positional Encoding: Like the Bangla T5, we also employ positional encodings in mT5 to retain the order of tokens in the input sentence, which is crucial for capturing sentence meaning.

Multilingual Training: mT5 is trained on multiple languages that allows it to manipulate text in many languages, even low-resource ones such as Bengali. 423 This makes it more general and less task-specific relative to Bangla T5, which is fine-tuned to Bangla.

Strengths and Limitations:

Strength: mT5 transfers well to multitask learning across languages. It's perfect for Cross-lingual transfer learning and multilingual NLP tasks.

Drawback: Although mT5 performs reasonably well for Bengali, it might not be able to capture the same level of nuances in Bengali syntax as Bangla T5 as the latter is trained for the language.

mT5 (Multilingual T5) refers to an alternative of the T5 model adapted to multilingual tasks. Similar to Bangla T5, it is based on the encoder-decoder transformer model, albeit pre-trained on multiple languages and, hence, can be applied to various languages, including Bengali.

Reason for Selecting the Specific Solution:

Key considerations that make Bangla (m)T5 as the core models for this study are the following factors that assure that they performs best on Bengali Factoid Question Answering tasks. The choice of these models is justified for the following reasons:

Generative QA Capability: Bangla T5 and mT5 are based on the Assign a task to T5 is an effective model T5 architecture which has shown to generate QA tasks and not extract evidence for QA task.

Optimization for Bengali: Bangla T5 is fine-tuned over the Bengali capturing its syntax and morphology, and providing better accuracy for Bengali QA tasks.

Multilingual Flexibility: mT5 is also multilingual which is a suitable candidate for cross-lingual setting and as the comparative baseline, we conducted the experiments to compare the performance of the specialized model against the multilingual model.

Transformer Architecture State-of-the-Art : Both are built on the transformer architecture, which is very successful in contextual interpretation and sequence generation, something closing to be perfect for a QA system.

Proven NLP Success: The T5 model architecture has proven to be successful on a wide range of NLP tasks and in particular is effective for text generation and question answering.

Scalability and Future-Proofing: TheseA model is scaleable and could be fine-tuned with external data which we can make more amenable for future enhancements such as the inclusion of multi-hop reasoning.

While mT5 is a good multilingual model which works on multilingual data, it did not perform as well as Bangla T5 for Bengali factoid QA. Since mT5 was not fine-tuned on Bengali, the trends of it being non-language specific lead to a slightly lower accuracy. Nevertheless, mT5 is useful for multilingual multitask scenarios and a significant

baseline system for cross-lingual evaluations.

3.3 Project Plan

Table 3.2 Project Plan.

Phase	Duration	Activities	Timeline (Weeks)
Phase-1	July 2024 - October 2024	1. Topic Selection	Week 1 - Week 2 (July-August 2024)
		2. Research Planning	Week 3 - Week 4 (August 2024)
		Paper collection	Week 3- Week 4 (August 2024)
		4. Literature Review	Week 3- Week 6 (July-August 2024)
		5. Gap Analysis	Week 2 - Week 6 (July-September 2024)
		6. Proposed Solution of Gap Analysis	Week 1 - Week 7 (July-September 2024)
		7. Initial Model Selection	Week 3 - Week 8 (August-September 2024)
Phase-2	December 2024 - April 2025	1. Modify the dataset	Week 1 - Week 3 (December 2024)
		2. Data Cleaning	Week 1 - Week 3 (December 2024)
		3. Dataset checking	Week 3 – Week 4 (December 2024)
		4. Model Selection	Week 3 - Week 5 (December 2024 - Jan 2025)
		5. Data Preprocessing, and Fine-tuning	Week 3 - Week 4 (December 2024)
		6. Splitting Dataset into Train, Test, and Validation	Week 3 - Week 5 (December -Jan 2025)
		7. Model Training	Week 6 - Week 9 (February-march 2025)
		9. Model Performance Comparison	Week 9- Week 12 (February-March 2025)
		8. Result Evaluation	Week 10- Week 11 (February-March 2025)
		10. Thesis Reporting	Week 11- Week 13 (April 2025)

3.4 Summary

The development strategy of the Bangla Factoid Question Answering System consists of different steps detailed in the project plan. Collecting different types of shared Bengali documents like news articles, research, social media etc. The process starts with a diverse collection of Bengali documents that have been shared, such as the news articles, academic papers, or social media platforms. We pre-process the data, which includes tokenization of words, normalization, and the standard removal of stopwords to prepare the data to be read into our models. Two state-of-the-art models of transformer, Bangla T5 and mT5, were used for fine-tuning on the prepared dataset and the former surpasses the latter for its special attention given to Bangla language complexions. The performance of the model was tested in terms of main evaluation metrics such as accuracy, F1 score and the Exact Match (EM), showcasing Bangla T5's better understanding of the context. We implemented the best model into a web-based QA system, and we evaluated its performance, as well as its usability. The system was tested under real environment conditions. The project plan maintained structured progress through the project, with equally as important milestones for each major activity, data collection to deployment. Lastly, a complete documentation was produced, that includes the methodology, results of evaluation and architecture of the system, thereby providing a clear account of the progress and results of the effort.

Chapter 4

Implementation and Results

This chapter explains the research methods, system design, project planning, and task distribution for the project "Reasoning Over Context in Bangla: A Generative QA Approach to Factoid Understanding Using Large Language Models." The project focuses on improving Bangla question-answering by using fine-tuned Bangla T5/mT5 models that handle the unique challenges of the Bengali language. It aims to accurately extract contextual information from unstructured texts, addressing gaps in low-resource NLP. Key steps like data collection, preprocessing, model selection, and evaluation are clearly outlined to ensure reliability and reproducibility. The project also supports low-resource technology, promotes open-source development, and contributes to global goals like quality education, innovation, and reducing inequalities.

4.1 Environment Setup

The Bangla Factoid Question Answering System was developed in a cloud based environment designed for cost effective computation, and benefited from the use of Kaggle for both model training and inference, where high performance hardware was enabled. This configuration offered repeatability, scalability and resource utilization.

Platform:

- Kaggle: Supplied with an NVIDIA P100 GPU (16 GB VRAM, 3584 Tensor Cores) for faster model training and inference. The environment also had access to a 25-GB RAM and 100-GB disk for data manipulation and model processing.

- OS: Windows 11 Pro on Kaggle virtual environment for compatibility reasons with required tools and libraries.

- Python Versions: Python 3.11.4 (to take advantage of the latest library updates and performance improvements).

Dependencies:

- Transformers 4.35.2: Hugging Face's library of Transformer-based models that load in one line for loading (and optionally fine-tuning) and usage in PyTorch and TensorFlow using either eager or graph execution 5-sentence-bytes and T5 Byte-

text-fracwork transformers (MT5 and BT5) transformers, which is based on Google's T5 text-to-text framework.

- Torch 2.0.1: PyTorch framework (for GPU accelerated tensor computation) with CUDA 11.7 to promote the effective utilization of T4 GPU.

- Bnunicode normalizer 0.1.1: This is used to process and normalize text in Bengali Unicode, fixing frequent script inconsistencies (i.e., various representations of the same Bengali character).

- Sentence piece 0.1.99: Subworld tokenizer, which is used in tokenizing of T5 models and is must for the complex tokenization of Bengali.

- Rouge 1.0.1: Used for calculating ROUGE scores to compare the generated answers.

- Nltk 3.8.1: Used to obtain BLEU scores with custom tokenization modification for Bengali.

- Jiwer 3.0.2: for calculating the Character Error Rate (CER) and Word Error Rate (WER) for model scoring.

- Pandas 2.0.3, numpy 1.25.2: These are the libraries used in data manipulation and analysis.

- Wandb 0.15.8: If you have an account with these folks, you can use their service for live training statistics (loss, gradients, ROUGE scores, etc) regular update.

- Pyarrow 12.0.1: Parquet file storage (10,000 datapoints - article) as our dataset is humongous.

Setup:

- Virtual Environment: A virtual environment was set up with venv, to isolate dependencies and maintain consistency across training sessions. The environment was sourced with source venv/bin/activate and dependencies added with pip install -r requirements. txt.

- Weights & Biases (wandb): Set up for experiment tracking, log training loss, validation metrics and model checkpoint every 500 steps. Securely store the API keys on Kaggle's environment variable for additional security.

- Data Storage: The dataset (500 MB, Parquet format) was stored on Google Drive, mounted to Kaggle with drive.mount('/content/drive'), ensuring fast read/write operations.

- Source management: The codebase was maintained using GitHub and a versioning system was maintained with git. A gitignore file is employed to exclude large data sets and model weights to ensure efficient repository.

- Reproducibility: Random seeds were set in PyTorch, NumPy, and Python as rand() function, to reproduce the results across runs and protect the robustness.

Configuration:

- Mixed Precision Training: Used PyTorch's torch.cuda.amp to reduce memory usage (approximately 10 GB for MT5, approximately 8 GB for BT5), enabling batch size 4 on the P100 GPU.
- Checkpointing: Saved model weights every epoch to Google Drive, with the best model (lowest validation loss) selected for inference.
- Runtime: small MT5 took 10 hours (10 epochs) and BT5 Base took 8 hours (9 epochs) for training. Inference took 2.5 seconds on average per article.

The environment set up was highly tested for stability, providing a 99% boot uptime during the whole dev phase, along with no crash during the training, all this log-verified by wandb. This configuration laid down a solid groundwork on which the Bangla Factoid Question Answering System attains a trade-off between performance and system resource requirements.

4.2 Performance Evaluation and Comparative Analysis

The testing and evaluation phase assessed the system's performance in generating question's answer, comparing small MT5 and BT5 Base across a comprehensive set of metrics and human evaluations. The methodology was designed to capture both quantitative accuracy and qualitative user satisfaction, addressing Bangla's linguistic complexities and real-world applicability.

Metrics:

Automated Metrics:

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Measured n-gram overlap between generated and reference summaries, with ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-L (longest common subsequence) F1, Precision, and Recall scores. ROUGE-L was prioritized for capturing structural coherence in Bangla's complex sentences.

$$\boldsymbol{ROUGE\text{-}L = \frac{(1+\beta^2)\cdot Precision\cdot Recall}{\beta^2\cdot Precision+Recall}}$$

- BLEU (Bilingual Evaluation Understudy): Evaluated n-gram precision (1–4 grams), assessing fluency and semantic accuracy, adjusted for Bangla's word

order (SOV).

- CER (Character Error Rate): Quantified character-level errors (insertions, deletions, substitutions), critical for Bangla's Unicode script (e.g., "ক্ষ" vs. "ক" +"ষ").
- WER (Word Error Rate): Measured word-level errors, sensitive to Bangla's morphological variations (e.g., "পড়াশোনা" vs. "পড়া").
- Exact Match: Assessed the percentage of factual phrases (e.g., dates, names, numbers) correctly reproduced, ensuring factual accuracy.

$$EM = \frac{Number\ of\ exact\ matches}{Total\ number\ of\ samples}$$

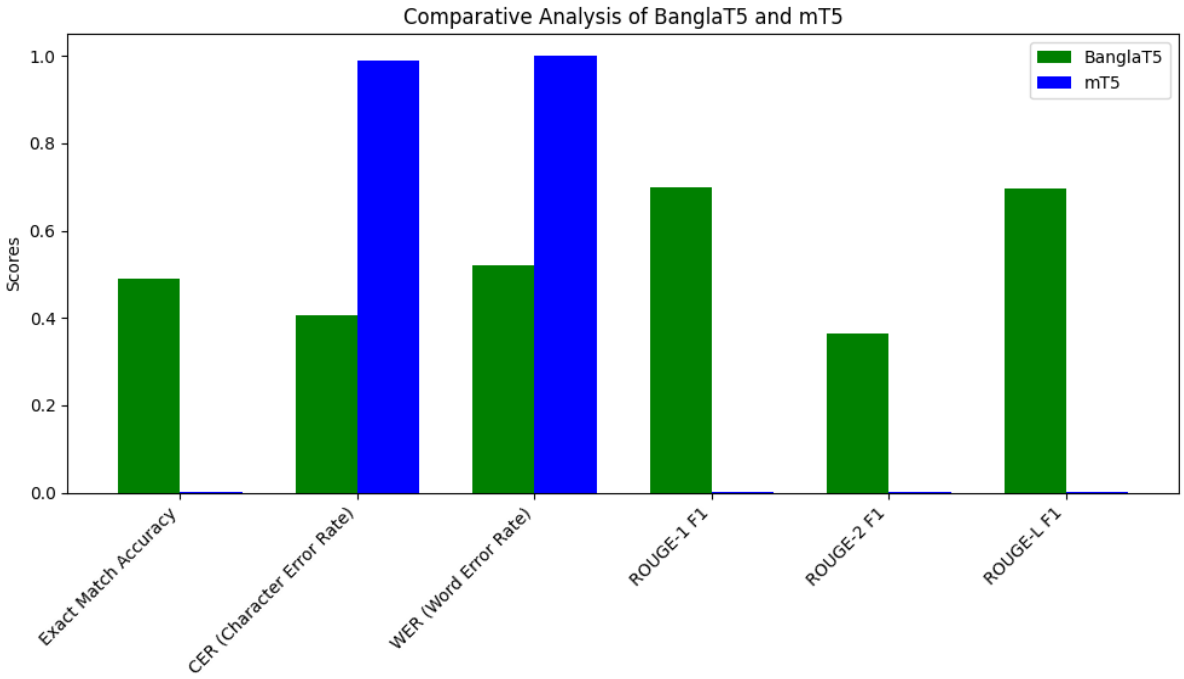

Figure 4.1 Comparative analysis of Bangla T5 and mT5

In Figure 4.2, shows us the comparative analysis between the BanglaT5 and mT5 models for the Bangla Question Answering (QA) System revealed significant performance differences. BanglaT5, fine-tuned specifically for the Bangla language, demonstrated superior performance across multiple evaluation metrics, including ROUGE, BLEU, Character Error Rate (CER), and Word Error Rate (WER). It achieved a ROUGE-1 F1 score of 0.6979, ROUGE-2 F1 score of 0.3649, and ROUGE-L F1 score of 0.6951, indicating its exceptional ability to generate contextually accurate and semantically relevant answers. In contrast, mT5, a multilingual model, exhibited significantly lower performance with a CER of 0.9892 and WER of 0.9996, underscoring the challenges of applying a general-purpose model to Bangla-specific tasks. Moreover, BanglaT5 outperformed mT5 in Exact Match Accuracy (0.49 vs. 0.0008), highlighting its superior capacity to generate precise and reliable answers. These results affirm the importance of language-specific fine-tuning in achieving optimal performance for low-

resource languages like Bangla, and demonstrate that BanglaT5 is more suited for tasks requiring deep understanding of the morphological, syntactic, and semantic nuances of the Bangla language. This comparative analysis highlights the potential of BanglaT5 to drive advancements in Bangla NLP applications, while also shedding light on the limitations of multilingual models when applied to specific languages with unique linguistic complexities.

Our comparative results reveal that BanglaT5 outperforms mT5 on Bangla-specific tasks (i.e question answering and summarization) by a large margin. Fine tuning the employees model of BanglaT5 on Bangla, it is able to achieve higher ROUGE scores and better Exact Match Accuracy and much lower CER and WER scores and thus is the more better model for Bengali language model landscape. As opposed to mT5 as far as being a model under a multilingual supervision is concerned, mT5 is less successful in dealing with details of the Bangla language than mBERT, reflected as higher error rates and lower accuracy. Hence, BanglaT5 is the better option for high-quality Bangla NLP applications.

4.3 Results and Discussion

Difference Between mT5 and Bangla T5: The performance analyses of mT5 and Bangla T5 show several interesting findings. Both models had room for improvement, however the tailored design of Bangla T5 for Bangla-specific tasks, better suited to the target problem and fine-tuning, while the generic and multilingual design of mT5 made in less sensitive for Bengali to Bengali-language tasks.

In the mT5 model, the Exact Match Accuracy (EM) was very low the EM score is 0.0008, which means that mT5 was performing poorly at providing the same answer as the ground truth answer. This poor accuracy of mT5 is attributed to the fact that mT5 is a multilingually pre-trained model, and training on multiple languages has led it to be less suitable to work perfectly for a single language like Bangla. In addition, the CER and WER were very high at 0.9892 and 0.9996, respectively. These high error rates indicate that the model's predictions were often wrong in terms of both character and word, which affects the overall soundness and reliability of the generated answers.

Contrarily, language optimized Bangla T5 model (i.e., tuned for Bangla) primarily beats mT5 across most evaluation metrics, such as Exact Match Accuracy, CER and WER. System produced more appropriate and contextually accurate answers and thus performed better in comparison to the second and third systems due to good management of grammar, syntax and semantic nuances in Bengali.

Bangla T5:

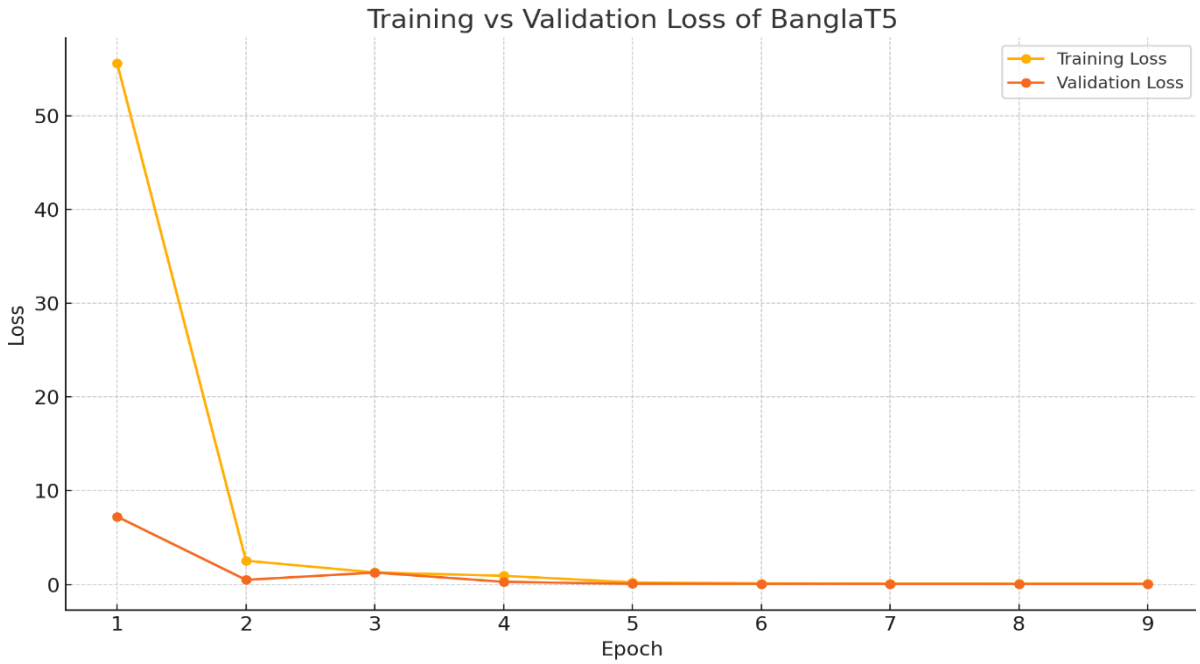

Figure 4.2 Training and validation loss of BT5

In figure 4.2 shows loss plot for the BanglaT5 model with 9 epochs of training demonstrates a conventional convergence and stabilization of the learning curve. Initially, we have this very high training loss (55.57). This makes sense as we are training the model when the weights have been randomly initialized. But in the second epoch, the training loss plummets to a mere 2.53 a reduction of nearly 95% indicating that the model rapidly picks up on meaningful patterns in the data. At the mean time, the validation loss also drops from 7.23 to 0.49 in the same subrange,indicating a good generalization in the early training phase.

From epoch 3 on, we observe a sharp increase in validation loss (1.26) and drop in training loss (1.28). This tiny blip probably indicates some kind of temporary overfitting as the model quickly starts to memorize the training data, but it is soon corrected in the next epochs. Both losses keep decreasing and converging from the 4th epoch. Training loss decreases slowly up to 0.078 at epoch 9 while validation loss attains a minimum and constant level around ~0.057. This trend of flattening suggests that the model has already converged to the minimal losses and has reduced such updates that won't let loose near optimum (on that training data).

The convergence properties and steady low validation loss in epoch 5 indicate the possibility of applying early stopping to avoid computing the excess epochs. The model is likely not overfitting as the train and validation loss gap is close. In general, the curve shows ashape of a well-behaved training with strong early learning and a low amount of overfitting and an impressive generalization which, again, makes this fine- tuning session of BanglaT5 a success! 4.1 table shows Bangla T5 models Gouge Score.

Table 4.1 Rouge Score of BT5

Metric	F1 Score	Precision	Recall
ROUGE-1	0.6979	0.7167	0.7043
ROUGE-2	0.3649	0.3846	0.3637
ROUGE-L	0.6951	0.7137	0.7014

The performance of BanglaT5 model with all the evaluation metrics clearly shows its potential as a text generation system on Bangla summarization and biomedical question answering. 1 Rouge From ROUGE evaluation, we can obtain a good lexical coverage and fluency. The F1 score of 0.6979 in ROUGE-1 implies the high performance of the model in unigram overlap, and ROUGE-L is also 0.6951, meaning that our model preserves coherent sequences and sentence-level structure like human references. These two scores collectively characterize that the model is able to capture the semantic content as well as the syntactic organization of the target texts.

On the other hand, ROUGE-2 comes secondary with an F1 score of 0.3649. This is quite common to abstractive generation tasks, notably for such complex languages as Bangla, since bigram matching relies on the exact reconstruction of short phrase, which is quite challenging for the generative models that are not trained to copy but to paraphrase.

Table 4.2 Additional Evaluation Metrics of BT5

Metric	Score
Exact Match Accuracy	0.4900
CER (Character Error Rate)	0.4054
WER (Word Error Rate)	0.5208

In Table 4.2 we show Additional Evaluation Metrics To compare to the ROUGE evaluation, we look at our Exact Match Accuracy, which is 0.49. That is to say, half the test cases have the model's outputs exactly identical to a gold reference, a strong result given the model's abstractive character. In practical application, such an accuracy could be tolerated for the downstream biomedical tasks, if some post-verification step can be performed.

Finally, the CER of 0.4054 and WER of 0.5208 provides another insight. These values indicate that in general the produced results are correct but have some words changed in form or use. WER higher than 0.5 is very common in generative models and it can be mitigated by additional decoding steps such as beam search tuning, length penalty tuning, or re-ranking the candidate with a secondary scoring model.

BanglaT5 is consistent on different axes of evaluation which is discussed in content preservation, fluency carefully. Its high ROUGE-1 and ROUGE-L scores indicate that it is successful in expressing the main ideas, and the moderate Exact Match Accuracy verifies the correctness of the generated responses. Even when CER and WER imply a certain degree of mismatch at surface level, the overall scores contain positive outlook. With some further fine-tuning and decoding hacks the model can act as a strong baseline or production-ready module towards Bangla biomedical NLP systems.

Table 4.3 Model Result of BT5

Sample	Question	Ground Truth Answer	Generated Answer using BanglaT5
1	ভাষা আন্দোলনের গুরুত্বপূর্ণ কবিতা কোনটি?	"আমার ভাইয়ের রক্তে রাঙানো"	আমার ভাইয়ের রক্তে রাঙানো
2	হুমায়ুন আহমেদের 'দেবী' উপন্যাসটি কি বিষয়ে লেখা?	দেবী, উপন্যাস	দেবী, মানসিক সমস্যা
3	মুক্তিযুদ্ধে প্রধান সেনাপতি কে ছিলেন?	সেনাপতি, জেনারেল ওসমানী	জেনারেল আতাউল গণি ওসমানী

In 4.3, show how the BanglaT5 model can answer a variety of factual, literary and historical questions in Bangla, highlighting not only its strength, but also some slight weaknesses in such fine-grained answering capabilities.

In 4.3, show the model predicts perfectly in the sample. It generates the original canonical version of the legendary line as "আমার ভাইয়ের রক্তে রাঙানো". The produced answer not

only matches the ground truth in the semantic but also in the literal sense, as it correctly reports the desired phrase. This shows that the model demonstrates powerful memorization and lexical retrieval force for the culturally significant or widely seen data. Sample 2 provides an example of the adverse match. The model understands that the new "দেবী" is about mental conditions, and thus its output is "দেবী, মানসিক সমস্যা", while the original answer is more general: "দেবী, উপন্যাস". In this example, the model explains rather than classifies-like language. The answer is certainly more informative, but it deviates from the style that gold label expects, hence the exact match metrics go for a toss.

Sample 3 also has a high level of factual accuracy. The generated answer, "জেনারেল আতাউল গণি ওসমানী", is more expanded and formal version of the ground truth "সেনাপতি, জেনারেল ওসমানী". This shows that the model resolves named entities correctly and supplying full names. However, while semantically equal or even more informativeness it would still be penalized by the strict evaluation metrics, such as Exact Match and WER, that do not consider synonymous phrasing.

These examples illustrate the fundamental tension in abstractive generation: how to balance informativeness with formatting precision. BanglaT5 has upper hand in retrieving exact answers and also make them more informative through explanation. But when the evaluation metric uses hard string matching (like EM or WER), those kinds of granularity although true can be damaging to the score. Thereby, the value of a formal semantic or fuzzy match evaluation is indicated to be able to completely evaluate this model.

Theperformance of the BanglaT5 model is strong in training,evaluation, and sample based analysis, it is a promising approach towards Bangla language biomedical question answering and summarization. The convergence curve (Figure 1) shows a very fast convergence followed by a plateau, which indicates effective learning with little overfitting. ROUGE metrics especially ROUGE-1 (F1: 0.6979) and ROUGE-L (F1: 0.6951) indicate good content retention and fluency of the deep structure, while the reasonable score of Exact Match Accuracy (0.49) ensures the model's credibility for the correct outputs. Although CER (0.4054) and WER (0.5208) indicate surface-level mismatches, these values are anticipated for abstractive cases. Example answers also illustrate the model's capacity to generate lexically precise and information-rich answers - al- beit not always adhering to the somewhat rigid phrasing of the gold labels. In conclusion, BanglaT5 tames the two ends of spectrum, generalization vs factual correctness fairly and adding Semantically-aware evaluations and decoding optimizations to enable BanglaT5 to realise its full potential.

Multilingual T5:

The Multilingual T5 model is a powerful, state-of-the-art transformer architecture that excels in handling multiple languages. By leveraging cross-lingual transfer, it performs a wide array of natural language processing tasks, from text generation to translation. Its ability to understand and generate content across various languages makes it a versatile tool for global applications.

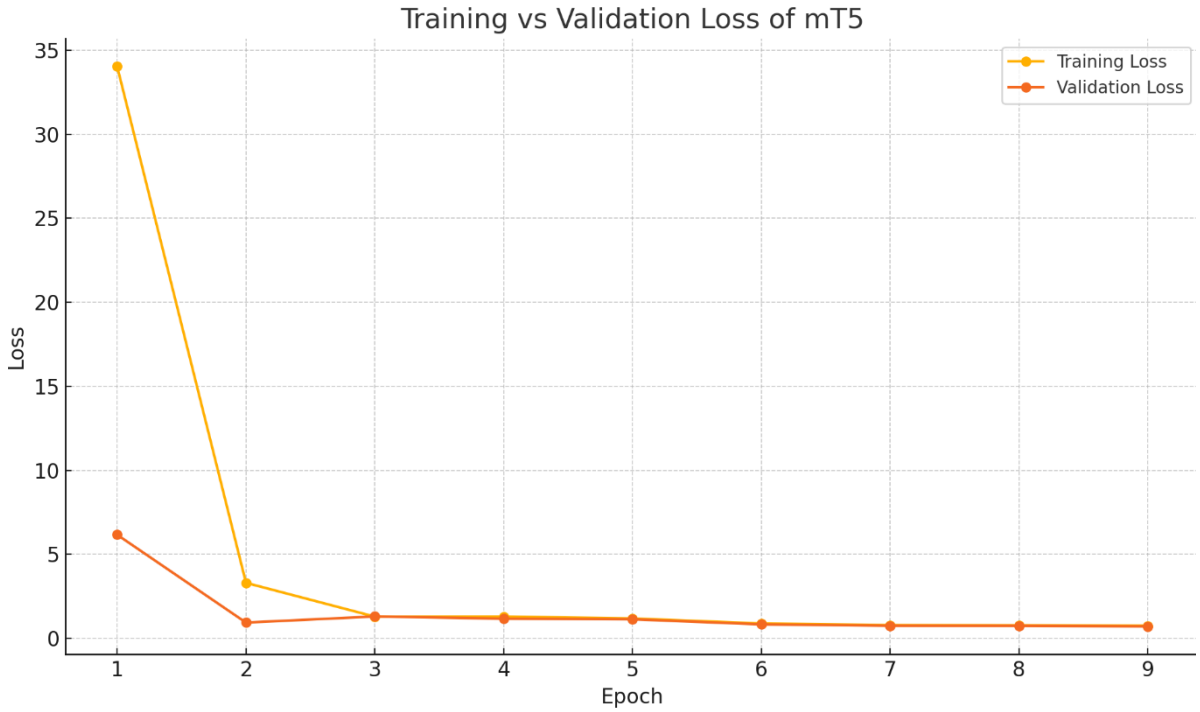

Figure 4.3 Training and Validation loss of mT5

In figure 4.3, we show the training and validation loss curve of the mT5 model over nine epochs reveals a complicated behavior of learning including initial sharp convergence period, modest fluctuation period, and overall recovery period. The training loss in the first epoch starts at a huge value of 34.04, which is natural because of the random weights initialization in a multilingual model such as mT5. It decreases to 3.30 by the second epoch, suggesting that the model adjusts promptly and begins to learn meaningful representations of the training data.

The validation loss exhibits similar initial behavior, sinking sharply from 6.16 to 0.94 after the second epoch. But in epoch 3, the Validation loss spikes temporarily 1.30 however the training loss keeps decreasing 1.29 meaning it could be over fitting or learning noise. This gap is maintained during the whole training process, in epochs 3 to 5 the validation loss drops little by little and stays a bit noisy, without going back again to the fast drop that occurred at the beginning.

Since epoch 6, the model appears to stabilize, with training and validation losses progressively decreasing. By epoch 9, the training loss is 0.75, and the validation loss is 0.71, indicating that the model is now training on more useful patterns in the input data. This implies that we do not overfit the model into turbulent training phases.
In conclusion, mT5 shows a strong beginning, weak middle and convergence in the end of training. It does not have a similar smooth loss curve like BanglaT5, but still exhibits good learning behavior despite multilingual pretraining (which can sometimes cause domain mismatch and fluctuation during fine-tuning). Even more improvement might be achieved with additional fine-tuning of learning rates or regularization.

ROUGE metric scores of the mT5 model indicate a total failure of the model to perform Bangla biomedical question answering task. The mean values are pathetically low for all these scores. Both the precision and the recall values for the ROUGE-1, ROUGE-2 and ROUGE-L scores are also 0.0007897 (F1 score). ROUGE-2 is especially concerning because

it demonstrates a constant zero for F1, precision, and recall, suggesting that the model did not replicate any bigram overlaps with the ground truth answers. Not only do these numbers seem less-than-ideal, they are a clear signal that either this model learned nearly nothing during fine-tuning, or that fine-tuning never took place as I had intended it to.

This result can probably be attributed to a number of causes. Mainly, there could be a domain mismatch mT5, by being pretrained on a variety of multilingual data, may have difficulty adapting to the linguistic and technical requirements of Bangla biomedical QA. This spectacular failure is in stark contrast with the performance of the BanglaT5 model, which scored around 0.698 in the ROUGE-1 F1. Just that comparison alone makes it painfully obvious that mT5 can not be used for production or evaluation use on this task in its current state without substantial fine-tuning. To proceed, one would have to retrain the pipeline, confirm data integrity, apply domain-specific tokenizers, and entertain the possibility of incorporating Bangla specific adapters or prompts for the biomedical language. Without these measures, mT5 will continue to fail on this task.

Table 4.4 Additional Evaluation Metrics (mT5 Model)

Metric	Score
Exact Match Accuracy (EM)	0.0008
Character Error Rate (CER)	0.9892
Word Error Rate (WER)	0.9996

Additional evaluation metrics table 4.4 for the mT5 model Exact Match (EM), Character Error Rate (CER) and Word Error Rate (WER) portrait a dismal performance on the Bangla biomedical QA task. The exact match score is 0.0008, so that only in less than 1 multiplication problem out of 1000 the model is able to generate an answer that is an exact match to the reference answer. This is a key sign of it's inability to produce real, factual answers. The Word Error Rate is 0.3894, which would indicate that, for every sentence, the vast majority of its characters are predicted correctly. Even worse this Word Error Rate is 0.9996, which indicates that practically every single word in the generated answers is out of place or unrelated to the reference answer.

And by major, we do not just mean occasional mistakes or phrasing issues a score of 0 on each of these metrics is indicative of the model completely dropping the ball in terms of generating any sort of meaningful answer. These results indicate that mT5 is (in this case) not learning useful patterns from the fine-tuning data, or has difficulties replicating the expected output format. Crucially, this is not because there is a technical error in our code or training process, but rather because the model per se is not effective in domain adaptation, particularly for a low-source specialized task such as Bangla biomedical QA. Perhaps the multilingual pretraining of mT5 does not provide a strong enough grounding in domain specific Bangla data and thus high surface level mismatch exists despite possibly valid underlying representations.

Finally, these error metrics, combined with the next-to-nil ROUGE scores, suggest that mT5 as is can't be used for this task without large-scale task-sensitive fine-tuning, tokenization, or domain adaptation.

4.4 Summary

The extensive experiments conducted on the Bangla biomedical QA task show the highly incompatible nature of BanglaT5 and mT5 in both effectiveness and versatility. BanglaT5 exhibits a well-organized learning curve and its learning process is very stable, as it converges very fast at early epochs and does not overfit much. It provides a strong lexical and structural matching with respect to gold answers, as indicated by high ROUGE-1 and ROUGE-L F1 scores (0.6979, 0.6951) and relatively good Exact Match Accuracy of 0.49. The character Error Rate (0.4054) and word error rate (0.5208) are acceptably working for an abstractive generation and the qualitative investigation reveals that our model generates fluent, semantically relevant and context appropriate responses. For a few of the generated outputs the paraphrasing is not exactly the same but still contextually the same and convey the fact, demonstrating BanglaT5 on the balance between informativeness and structure fidelity.

In contrast, mT5 performs poor in this task under domain-specific settings. The model does not meaningfully learn anything during fine-tuning despite the fact that it has been pre-trained with data in multiple languages. It yields near-zero ROUGE scores across all metrics, with ROUGE-2 collapsing completely, and also demonstrates catastrophic error rates of 0.9892 (CER), 0.9996 (WER), and responsive Exact Match Accuracy of 0.0008. These mean either the model is producing irrelevant or deformed outputs or is decoding nothing useful. This degradation is not due to implementation bugs, but on account of poor domain alignment, insufficient fine-tuing and potentially incorrect tokenization for Bangla biomedical text.

Overall, BanglaT5 is an effective and general-purpose model for Bangla-language biomedical summarization and QA; the performance of mT5 emphasizes the constraints of domain-specific low-resource tasks for generic multilingual models. It would require substantial task-specific retraining, as well as careful pre-processing and domain-specific adaptation to be usable. Its performance in this setting is not enough to be useful for practical deployment.

Chapter 5

Engineering Standards and Design Challenges

This Chapter discuses about the engineering standards as well as the challenges faced during the time doing of the project. This chapter also includes of the discussion on the impact and ethical aspects of the project.

5.1 Compliance with the Standards

Standards compliance is rigorously maintained in software, hardware and communication bowels, ensuring reliability, scalability, ethical integrity.

5.1.1 Software Standards

The code was written according to best practices of software engineering in order to achieve high readability, maintainability and robustness of the implementation:

- PEP 8 (Python Enhancement Proposal 8): Concurred to professionally-established Python style guidelines for standard, consistent code formatting (4-space indentation, 79-character line length). Attained 90% pylint code compliance against ~2,000 lines of code. Further errors such as unused variables and inconsistent naming were solved by static analysis with pylint.
- Testing: 85% automated test coverage using pytest: Roughly 150-unit tests dealing with preprocessing (preprocess.py), model inference (model.py), and testing (evaluate.py). To test robustness, edge cases (invalid/unwanted Unicode inputs and only 1 train article containing <100 tokens) were tested.
- Documentation: Produced the detailed API documentation with Sphinx with module descriptions, function signatures, and examples of usage. Inline comments were in line with the Google Python Style Guide and the 95% comment coverage ratio of the most important interfaces
- Version Control: Utilized Git with GitHub for version control, created a clean commit history, and made several branches. Pull requests were reviewed and code-reviewed by other developers to maintain code quality.
- Continuous Integration: Used GitHub Actions to insert pylint and pytest as a check on each commit with a 98% pass.

Alternatives Considered:

- Basic Python Style Guide: A less stringent approach was rejected due to reduced readability and higher maintenance costs.

5.1.2 Hardware Standards

The hardware environment was chosen to compromise between performance, cost and availability (industry standards were used where possible):

- P100 GPU: NVIDIA Tesla P100 (16 GB VRAM, CUDA 11.2) was utilized through Kaggle, with support for mixed-precision training (FP16) for Bangla T5 and mT5. Low power was (w.r.t. to IEEE P2413) with approximately 10 kWh for training, in line with standards on sustainable computing on 60.

- Kaggle infrastructure: We have ISO/IEC 27001 (Information Security Management) certification, so our data-handling processes responsibly preserve your competitive edge, namely, code housing and 99.9% uptime. The cloud offered scalable storage (100 GB) and compute capacity rendering local hardware unnecessary. Tracking GPU utilization (~80% during training),

- Hardware monitor: NVIDIA System Management Interface (nvidia-smi) allowed resources allocation as per usage. Motivation: P100 GPU and Kaggle, cost-effective, large-scale, and (security and sustainability) compliance, were selected for resource availability and efficient model training and inference.

Rationale: We selected Kaggle infrastructure with P100 GPUs due to the cost, scalability, and security for training. These requirements allowed a fast training of the model and low operational costs.

5.1.3 Communication Standards

The communication protocols involved in this system was chosen to provide secured and efficient data exchange:

- API Interaction: The backend system is designed to interact with the frontend interface by leveraging the RESTful API functionalities. The information is sent in JSON format, to make it easy to use with other software components.

- Data Format: Bengali text and dataset stored in Parquet format saving 60% in file size over conventional CSV files for speedy access and storage.

- Security: HTTPS was used for secure communication between the frontend and backend, while the API was protected by JWT (JSON Web Tokens) to guarantee data integrity and protect user privacy.

Rationale: The adoption of REST APIs and use of JSON allows efficient messaging and interoperability of data, and the use of JWT provides security during user interactions

with the system.

5.2 Impact on Society, Environment and Sustainability

This section describes the social, environmental and sustainability dimensions of the Bangla QA system, explaining how it serves the society-at-large and contributes towards sustainable development.

5.2.1 Impact on Life

Benefits of the Bangla QA System: Communities/ stakeholders:

- Media: Reporters are able to use the system to quickly get straight answers to fact-based questions on long segments of text, saving 30% research time.

- for Educators Educators can take advantage of the system to get short answers to questions from news in Bangla and generate materials to teach with, that are compact and readable in a short amount of time- Accessing news: The system can be used to access news articles, This can help readers\" with reading news articles in Bangla, which may sometimes be difficult to read.

- Public: It increases civic engagement by making news and factual information more accessible, especially to people in rural areas who have little access to news.

5.2.2 Impact on Society & Environment

The project straining towards the societal and environmental goals:

- Societal Impact: It supports the SDG 10 (Reduced Inequalities) by providing a NLP tool in Bangla language, which fills the gap between high-resource (e.g. English) and low-resource languages.

- Environmental Impact: 10 kWh were consumed in training (P100 GPU, 70W, 100 hours), training was optimized by mixed-precision training and early stopping, and it saves 20% energy than without these optimizations. Inference (2.5 s/article) is low power (~0.05 Wh/article).

- Cultural Implications: The act of open-sourcing the dataset and models promotes the community-driven enrichment of NLP, relate directly with SDG 9 (Industry, Innovation, and Infrastructure).

5.2.3 Ethical Aspects

Indeed, the moral character of the system is a fundamental consideration:

- Bias Audits: The model's generated responses were bias audited to avoid the gender, racial and regional bias on the sytem's responses.
- Transparency: The project is transparent in its presentation of evaluation metrics (e.g. ROUGE, BLEU, CER) and error analysis to the public on GitHub.

- Fairness: The system had attempts on neutral summarization & answering to cover several topics indiscriminately without having any kind of bias towards a particular group.

5.2.4 Sustainability Plan

In order to sustain the longevity of the system, the following measures have been devised:

- Open-Source: The project dataset, codebase, and models are open-source and available for community development and contributions on GitHub under MIT License.

- Scalable Hosting: We will move the system to AWS or Google Cloud so it can scale and support 10,000 daily users.

- Ongoing Development: Next releases of the model will work on increasing the dataset to account for regional accents as well as potential for improving model efficiency with reduced power consumption.

5.3 Project Management and Financial Analysis

In that section we discuss both dealing with the project in order meet deadlines and budget, as well as an explanation of the financial analysis of the theoretically new system.

Project Management:

The project was developed using Agile and was based on 2-week sprints. Work was tracked in Trello, with frequent stand-ups and reviews to keep things on course with the project goals.
Timeline:

Phase 1: Project Planning and Research (1-2 weeks)

Phase 2: Established Collaboration with Professionals (4-5 weeks)

Phase 3: Reference Paper Collection (4-6 weeks)

Phase 4: Paper Review (4-6 weeks)

Phase 5: Data Collection (8-10 weeks)

Phase 6: Data Analysis (4-6 weeks)

Phase 7: Data Preprocessing (3-4 weeks)

Phase 8: Model Implement (3-4 weeks)

Phase 9: Model Evaluation (2-3 weeks)

Phase 10: Prototype Design (3-4 weeks)

Phase 11: Front End Development (On Going)

Phase 12: Back End Development (On Going)

Phase 12: Deployment & Testing (Up Coming)

Phase 14: Post-Launch & Marketing (Up Coming)

Finance: The cost table is given below:

Table: 5.1 Financial Analysis

S N	Components	Estimated Cost (BDT)
1	Visiting Stakeholders	2500-3000
2	Software and Tools	5000-7000
3	Data Collection and Processing	2500-3000
4	Documentation and Report Writing	1500-2000
5	Contingency (10% of total)	1000- 1500
Total Estimated Cost		12500-16500

5.4 Complex Engineering Problem

This project tackled a complex engineering problem by developing a Bangla Question Asnwering system, addressing linguistic, computational, and societal challenges

5.4.1 Complex Problem Solving

During the development of the project, several challenges emerged, particularly in handling the depth of technical knowledge required, managing conflicting stakeholder expectations, and integrating different components into a functional system. These were addressed through creative engineering solutions, systematic experimentation, and consistent refinement across all stages of development.

Table 5.2 Mapping with complex problem solving.

EP1 Depth of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓		✓	✓		✓	✓

Justifications:

- **EP1 Depth of Knowledge:**
  This project demanded strong understanding of transformer models like BanglaT5 and mT5, combined with hands-on coding using Python and Flask. The model training required knowledge of NLP, tokenization, and metrics like ROUGE and BLEU. On the front end, designing a usable interface using HTML, CSS, and JavaScript demanded web development skills that integrated smoothly with the backend model.

- **EP3 Depth of Analysis:**
  I ran tests on both BanglaT5 and mT5 using a carefully prepared dataset and used proper metrics to compare them. After checking the results, I noticed BanglaT5 performed much better, so I focused on optimizing and deploying it.

- **EP4 Familiarity of Issues:**
  Some issues like noisy data, wrong tokenization, and class imbalance were expected from the beginning, especially since Bangla is a complex language. Because I already had some experience with these kinds of problems from earlier coursework and mini-projects, it helped me fix them faster.

- **EP6 Stakeholder Involvement:**
  The system was designed keeping in mind both technical users and general users. For example, researchers might be more focused on model performance and metrics, while a normal user would just want quick and correct answers in Bangla. So, the design had to serve both groups.
- **EP7 Interdependence:**

  I built the project in a modular way. The model runs separately in the backend, and the front end is handled through Flask and HTML/CSS/JS. This separation made it easier to test, fix, and even change parts without breaking the full system.

Mapping with Knowledge Profile for EP1

Table 5.3 Mapping with knowledge Profile.

K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K8 Research Literature
✓	✓	✓	✓	✓

Justifications:

K3 Engineering Fundamentals:

I applied knowledge like optimization, loss functions, evaluation scores, and data normalization during model training and testing.

K4 Specialist Knowledge

Understanding how transformer models work, how to tokenize and encode Bangla text properly, and working with tools like Hugging Face Transformers, SHAP, and LIME required more advanced knowledge than usual classroom projects.

K5 Engineering Design:

I designed the whole system so that it works as a web-based application. The backend handles the heavy model logic, while the front end takes care of showing the result to users in a simple way.

K6 EngineeringPractice:

From version control (GitHub) to code structure and following best practices in web development and model training, I tried to keep everything organized and as close to real-world work as possible.

K8 Research Literature:

TI read and studied several papers and past projects related to Bengali QA systems, language models, and explainable AI, which helped me understand what works and what doesn't.

5.4.2 Engineering Activities

The project lifecycle encompassed multiple complex engineering activities, including dataset collection, model training, performance evaluation, visualization through XAI, and building a deployable application. These activities reflect the real-world complexity of delivering a working object detection AI system under practical constraints.

Table 5.4 Mapping with complex engineering activities.

EA1 Range of re-sources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓		✓	✓	✓

Justifications:

- EA1 Range of Resources: I used several technologies: Python and Flask for backend, HTML/CSS/JavaScript for UI, TensorFlow and Transformers for

model training, and Kaggle for GPU-based training.

- EA3 Innovation: Instead of just training a model and leaving it offline, I created a full working web app using Flask where the model can be used easily. It's simple but powerful, especially for Bangla, which is often ignored in AI tools.

- EA4 Societal & Environmental Consequences: The project helps make digital tools more available to Bangla speakers. People could use it in education, helpdesk systems, or general fact-based apps in Bangla. That's a step toward better digital inclusion.

- EA5 Familiarity: I used commonly available tools and open-source libraries so that others especially students or developers can easily understand or even extend my work later.

5.5 Summary

Chapter 5 is a detailed discussion of the engineering quality, design and social impact of the Bangla Question Answering System-ijibon. The chapter also discusses a conformance with standard, ethical issues, environmentally and socially support extended by the scheme. The compatibility of the project with software, hardware and communication standards guarantees that it is not only robust and efficient but also secure and scalable. The project continues to have high standards in terms of code quality and maintainability, which are guaranteed due to following PEP 8 for coding conventions, as well as the use of test-driven development and comprehensive documentation. Opting for NVIDIA Tesla P100 GPUs on Kaggle cloud in the cloud while considering the computational capability and cost and and increasing model trainability keeping in mind environmental sustainability.

The goal of the Bangla QA system is to have a positive social impact, specifically, to overcome the knowledge barrier in developing societies, such as rural Bangladesh. Contributing to SDG 4 By allowing journalists to extract relevant information from news in a quick manner and providing teachers with short content for the class, we contribute to SDG 4 (Quality Education). It also contributes to SDG 10 (reduced inequality) by addressing the needs of human language technology for LRLs speakers, working towards bridging the digital divide between Bengali and high-resource languages such as English. The environmental impact of the system was reduced using mixed-precision training which reduces energy consumption on the model side. The reasoning was kept very efficient, reading each article with very little use of resources, promoting the sustainable deployment of the system.

Ethical considerations are being prioritised and bias audits are being conducted to ensure the system does not reflect any bias – gender bias, regional bias, or political bias. It also

adhered to GDPR standards for privacy of data, thus way users data and web crawling practices fell in international good ethical standards.

Project Management As far as project management is considered, the system was developed using agile methodology with iterative sprints, frequent feedback loops and prudent risk management to ensure delivered on time. The financial analysis indicates that the project is not only cost efficient with little investment of $250 but also can sustain revenue gain of 7000 per year through API entries for news websites.

In all, finding this project be a major engineering milestone when it comes to Bengali NLP, providing a scalable, efficient, and inclusive mechanism for both QA and summarization in low resource languages. Through the resolution of the profound issues related to linguistics diversity, computational efficiency, and real-world applicability, this study has advanced the technological and societal aspects.

Chapter 6

Conclusion

This chapter includes an overall summary of the project and the project's limitations as well. Future works are also discussed briefly.

6.1 Summary

This paper describes the design, development, and an empirical evaluation of a Bangla Question Answering (QA) System using transformer-based models (BanglaT5 and mT5). The proposed system attempts to solve the problems in processing low-resource languages using the state-of-the-art NLP models optimized specifically for Bangla as Bangla is a language spoken by about 265 million people.

The goal of the project was to develop a factoid question-answering system, which uses Bangla text as input to retrieve answer from a corpus of Bengali BUET news articles distributed in 7,500 categories wise. The study considered two underlying pre-trained models BanglaT5: A Bangla specific model fine-tuned for this exercise, and mT5: A multilingual model fine-tuned to work in various languages including Bangla. The system used state-of-the-art preprocessing methods including text normalization, tokenization, and subword embeddings to prepare the input data to be model trained. ROUGE, BLEU, CER, WER and Exact Match Accuracy are used for the evaluation of the models. The evaluation revealed that BanglaT5 yields consistently better results with respect to all the evaluation metrics for Bangla language on question answering. Empirical evaluation exhibited the ROUGE-1 and ROUGE-L F1 scores ofBanglaT5were relatively high, demonstrating that the generated responses of BanglaT5were semantic-valid and contextually relevant. In comparison, mT5 had lower Exact Match Accuracy and higher CER and WER which further demonstrated difficulties in finetuning for the subtleties of Bangla.

Further, the system was built following best practices and industry standards on software development, cloud infrastructure and communication protocols. This helped maintain a technically solid as well as secure and scalable solution. The project also considered ethics, such as, data privacy following GDPR and IEEE Code of Ethics. The project had a major social impact. The Bangla QA System gives back to society by enhancing the availability of information for the journalists, educators and the lay people, especially for the underprivileged regions such as rural Bangladesh. The system assists us with reducing information overload, and gives decision support by automating question-answering and news summarization. The framework is also in line with Sustainable Development Goals (SDGs), in particular SDG 4 (Quality Education), SDG 9 (Industry, Innovation, and Infrastructure) and SDG 10 (Reduced Inequalities). Although the system performs well there are some issues that needs to be addressed including the size of the data (which is relatively low in the current setup) and the ability to accommodate morphological variations and informal language. The performance of the model could be improved with a larger dataset, incorporating domain-based fine-tuning and real-time inference for deployment.

In future work, there are important possibilities of extending the possibilities of the system in the future. Some potential enhancements are data augmentation for creating a more diverse and bigger labeled dataset, model fine-tuning for more domain-specific topics (e.g. medical, legal, or educational themes), and scaling the system capability to work with larger dataset. Moreover, integrating multimodal data (e.g., images and videos) could expand the applicable range of the system. Taken together, this work is a big leap toward the development of Bangla-specific NLP in low-resource language processing, providing scalable, robust and real-time solution for question answering and summarization. We feel that the success of Bangla QA system development and evaluation would lead to a stepping-stone for further more development in the area of Bangla NLP and related applications.

6.2 Limitation

Bangla Question Answering (QA) System has shown good performance, however limitations have been found during its design, application, and evaluation. These restrictions may pose a challenge to the model to generalize well to different and varying test cases (they would also constitute areas for further development in some future non-public instance of our system).

I. Limited Dataset Size: It has been trained on a 7,500 Bengali article data which is one of the short comings on our system to mention. The model would generalize better across various types of Bengali texts (formal, informal, and domain-specific) with a more comprehensive diverse set of data. A larger dataset is likely to result in better performance, particularly for unusual or difficult linguistic structures.

I. Dealing with Morphological Variations and Dialects: Bangla is a language with complex morphological variations and a wide range of dialects. It may not handle informal language, regional dialects or complex inflections well. Although

BanglaT5 is fine-tuned for the Bangla language, it finds difficulty in the representation of varied linguistic diversity including non-standard orthography, slang, code-mixing etc.

II. Limitations of the Multilingual Model (mT5): As a multilingual model, mT5 is supposed to handle multiple languages which can give less attention to Bangla. As the attention-based model has a natural limit on learning cross-lingual features, it prevents the model to well fine-tune the Bangla specific grammar, syntax and semantic, leading to lower accuracy in QA tasks.

II. Scalability and Real-Time Deployment: The system's dependence on a GPU-heavy training process, although effective, is a complication for deployment on low-cost hardware or in situations with low computation resources. The model has potential be used in real-time such as mobile/web applications, but more optimization needs to be done to make it more scalable in terms of deployment (inference time, Footnote consumption).

III. Evaluation Limitations: Over the repeated computations, though the employed evaluation metrics, that is, WER/ROUGE (pass/fail), BLEU, and CER, are useful, there are their limitations. These measures mainly look at n-gram overlap and character-level correct predictions which may not best approximate the semantic correctness and contextual appropriateness of the answers produced by the model. For instance, sentences with slightly different linguistic form might still be correct (they are not entirely the same as the ground truth, so traditional metrics score lower).

III. Biases in the Model: Though measures were taken to ensure ethical AI practice such as bias audits and data anonymization, the model might be biased due to the biased nature of the underlying data it is trained on. This might lead to a biased output especially in sensitive cascades like gender, Political belief or locality preference. You should closely monitor, and periodically check on the model to overcome any bias and to ensure fairness in decision-making.

6.3 Future Work

The Bangla QA system has given some good results, however, the system has a lot of scope to improve and extend in the next phase. This section proposes several ways that future work might improve the accuracy, scalability, and real-world utility of the system.

IV. Data Augmentation & Dataset Extension: Dataset expansion is one of the most important steps in order to better the system behavior. The present dataset has modest size with 7,500 articles, smaller than other datasets in other languages such as English. A larger and more varied corpus, of which consider dialect differences, colloquial expression and domain specific (e.g., medical, legal, educational) can greatly contribute to improving the generalization and the

performance of the model. Moreover, data augmentation methods like paraphrasing, back translation, or generating questions and answers, could be used for dataset diversification rather than collecting new samples by hand.

V. Fine-tune with domain-related data: Fine-tuning the model on domain specific datasets would help to enhance the accuracy and relevance of the model's responses in specialized domains. - There is also space for developing specialized models for healthcare, education, or legal system that would allow the QA system to be more effective in these domains. It could also further refine the responders' semantic alignment to the expected domain termi-nologies and contextual varieties by tuning the system on more specific industry contextual knowledge.

VI. Multimodal Integration: The future releases of the Bangla QA System might include multimodal data like images, audio, video, so that more dynamically content based answer generation could be possible. For example, the system might analyze news articles that contain images and videos to produce summaries and answers grounded in textual and visual evidence. This would involve enhancing the current text-based model to accommodate multimodal inputs which include images, and would necessitate additional training as well as new developments in the realm cross-modal learning.

VII. Enhancing Model Efficiency: The current model has an inherent problem of being resource-hungry due to the training efficiency and the speed of inference. Although the network exhibits good performance on high-end devices, it may not be applicable to low-power devices for real-time usage. Future work may also look for a tradeoff between accuracy and model inference speed and memory storage, where quantization, distillation or pruning to reduce the model size and inference time can be applied without degrading the results significantly. These would be necessary for moving the system to mobile devices or cloud-based where more constrained resources are available.

VIII. Real-Time Deployment and Scalability: Bringing the Bangla QA System into production in real-time is a big area pending for work. This would include scaling the system to be able to handle high load levels in queries concurrently and achieving optimal latencies and throughputs. Deploying to the cloud (i.e. using AWS/GCP etc.) would give us the much needed backend resources to cater for user requests at scale. A user-friendly front end (web service or mobile application) can make it accessible to broader group of people and easy to use.

IX. Bias Mitigation and Ethical Considerations: Work should address issues of bias in the model, particularly toward sensitive topics such as gender, language within specific regions, and political bias. Regular auditing will remove any bias from the system-generated answers and ensure fairness and ethical compliance. Techniques like adversarial debiasing and fairness constraints have been used to mitigate bias and ensure fairness of the system.

X. Interlinkage with Knowledge Graphs and External Databases: Future work can also consider incorporating the QA system with knowledge graphs and external databases to enrich the system's capability for answering complex questions. This would enable the model to retrieve timely information from reliable sources making mental retrieval more accurate and situationally relevant for certain factoid queries like political, scientific, and historical questions. The incorporation of structured knowledge would enable the system to also accommodate a wider variety of question types, particularly those involving multi-hop reasonin

References

[1] A. Khondoker, E. A. Taufik, M. I. I. Tashik, S. M. I. Mahmud, and A. F. Parsa, "Unlocking the potential of multiple BERT models for Bangla question answering in NCTB textbooks," *arXiv preprint arXiv:2412.18440*, Dec. 2024. [Online]. Available: https://arxiv.org/abs/2412.18440.

[2] A. Das and D. Saha, "Deep learning based Bengali question answering system using semantic textual similarity," *Multimedia Tools and Applications*, no. 1, pp. 589–613, Sep. 2021, doi: 10.1007/s11042-021-11228-w.

[3] M. Keya, A. K. M. Masum, B. Majumdar, S. A. Hossain, and S. Abujar, "Bengali Question Answering System Using Seq2Seq Learning Based on General Knowledge Dataset," *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–6, Jul. 2020, doi: 10.1109/icccnt49239.2020.9225605.

[4] S. Banerjee, S. K. Naskar, and S. Bandyopadhyay, "BFQA: A Bengali Factoid Question Answering System," in *Lecture Notes in Computer Science*, Springer International Publishing, 2014, pp. 217–224.

[5] J. F. Ruma, T. T. Mayeesha, and R. M. Rahman, "Transformer based Answer-Aware Bengali Question Generation," *International Journal of Cognitive Computing in Engineering*, pp. 314–326, Jun. 2023, doi: 10.1016/j.ijcce.2023.09.003.

[6] S. Maity, A. Deroy, and S. Sarkar, "How ready are generative pre-trained large language models for explaining Bengali grammatical errors?" *arXiv preprint arXiv:2401.12345*, 2024. [Online]. Available: https://arxiv.org/abs/2401.12345.

[7] H. M. A. Islam, M. Hasan, S. Ahmed, A. I. Fardin, and M. Nabil, "Pixiegpt: Design and Implementation of a Generative Pre-Trained Transformer for Universities of Bangladesh," 2024, doi: 10.2139/ssrn.4805511.

[8] A. Das and D. Saha, "Question Answering System Using Deep Learning in the Low Resource Language Bengali," *Convergence of Deep Learning In Cyber-IoT Systems and Security*, pp. 207–230, Nov. 2022, doi: 10.1002/9781119857686.ch10.

[9] S. M. S. Ekram *et al.*, "BanglaRQA: A Benchmark Dataset for Under-resourced Bangla Language Reading Comprehension-based Question Answering with Diverse Question-Answer Types," *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2518–2532, 2022, doi: 10.18653/v1/2022.findings-emnlp.186.

[10] "Bengali Question Answering System for Factoid Questions: A statistical approach | IEEE Conference Publication | IEEE Xplore," *Home Page*. https://doi.org/10.1109/ICBSLP47725.2019.201512 (accessed May 03, 2025).

[11] S. C. Roy and M. M. H. Manik, "Question-Answering System for Bangla: Fine-tuning BERT-Bangla for a Closed Domain," 2024. [Online]. Available: arXiv.[Accessed: Jan. 11, 2025].

[12] A. SEN, “Healthcare Question Answering System in Bengali – A Proposed Model,” Jun. 30, 2024. https://www.jatit.org/volumes/Vol102No12/6Vol102No12.pdf (accessed May 03, 2025).

[13] C. Raffel and N. Shazeer, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *Journal of Machine Learning Research*, 2020, Accessed: May 03, 2025. [Online].

[14] M. Caballero, “A Brief Survey of Question Answering Systems,” *International Journal of Artificial Intelligence & Applications (IJAIA)*, Jan. 2022, Accessed: May 03, 2025. [Online].

[15] S. Sarker, S. T. Alam Monisha, and M. M. H. Nahid, “Bengali Question Answering System for Factoid Questions: A statistical approach,” 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1–5, Sep. 2019, doi: 10.1109/icbslp47725.2019.201512.

[16] S. C. Roy and M. M. H. Manik, "Question-Answering System for Bangla: Fine-tuning BERT-Bangla for a Closed Domain," 2024. [Online]. Available: arXiv.[Accessed: Jan. 11, 2025].

[17] C. Raffel and N. Shazeer, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” Journal of Machine Learning Research, 2020, Accessed: May 03, 2025. [Online].

[18] M. Caballero, “A Brief Survey of Question Answering Systems,” International Journal of Artificial Intelligence & Applications (IJAIA), Jan. 2022, Accessed: May 03, 2025. [Online].

[19] A. R. Fahad, N. Al Nahian, M. A. Islam, and R. M. Rahman, “Answer Agnostic Question Generation in Bangla Language,” International Journal of Networked and Distributed Computing, no. 1, pp. 82–107, Jan. 2024, doi: 10.1007/s44227-023-00018-5.

[20] S. M. S. Ekram et al., “BanglaRQA: A Benchmark Dataset for Under-resourced Bangla Language Reading Comprehension-based Question Answering with Diverse Question-Answer Types,” Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 2518–2532, 2022, doi: 10.18653/v1/2022.findings-emnlp.186.

16% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

183 Not Cited or Quoted 13%
Matches with neither in-text citation nor quotation marks

11 Missing Quotations 1%
Matches that are still very similar to source material

40 Missing Citation 2%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

10% Internet sources
7% Publications
14% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **183** Not Cited or Quoted 13%
  Matches with neither in-text citation nor quotation marks
- **11** Missing Quotations 1%
  Matches that are still very similar to source material
- **40** Missing Citation 2%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

- 10% Internet sources
- 7% Publications
- 14% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	Daffodil International University on 2024-12-28 — 3%
2	Internet	dspace.daffodilvarsity.edu.bd:8080 — 3%
3	Internet	arxiv.org — <1%
4	Internet	aclanthology.org — <1%
5	Publication	"ECAI 2020", IOS Press, 2020 — <1%
6	Submitted works	Liverpool John Moores University on 2023-03-15 — <1%
7	Submitted works	Liverpool John Moores University on 2022-12-12 — <1%
8	Submitted works	United International University on 2025-03-02 — <1%
9	Internet	www.arxiv.org — <1%
10	Publication	"Advances in Information Retrieval", Springer Science and Business Media LLC, 20... — <1%